3L: The Southeast Asian Journal of English Language Studies – Vol 24(3): 26 – 42
http://doi.org/10.17576/3L-2018-2403-03

What Do Different Word Lists Reveal about the Lexical Features of a Specialised Language?

NOORLI KHAMIS
Pusat Bahasa dan Pembangunan Insan (PBPI)
Centre for Technopreneurship Development (CTeD)
Universiti Teknikal Malaysia Melaka (UTeM)
noorli@utem.edu.my

IMRAN-HO ABDULLAH
School of Language Studies and Linguistics
Universiti Kebangsaan Malaysia

ABSTRACT

Most corpus-based investigations capitalise on word list analyses: frequency, keyword, and key-keywords, in profiling the lexical features of a specialised language. Though the three word lists have been used in many corpus-based language studies, comparisons across these three types of word lists in characterising a specialised language has not been made to identify any salient information each word list can reveal about the target language. This paper provides comparisons of Engineering English using three types of word list: frequency, keyword and key-keyword lists. The purpose is to identify the lexical information that can be revealed by the groups of words listed according to each type of word lists. To conduct the analyses, a corpus of Engineering English (E^2C) is created. All the word lists from the corpus are extracted using the Wordsmith software. Next, further analyses on the distribution of the vocabulary components, namely function vs. content words, and word categories i.e. GSL, AWL and Others, are conducted on all the three word lists. The findings reveal that different word lists result in different ranges of words, and the analyses of the words reveal the distinct features of the specialised language at different levels. Given such differences, this study provides insights into which word lists are to be considered in a lexical study for language description purposes. Hence, this study further verifies the importance of corpus-based lexical investigations in providing empirical evidences for language description.

Keywords: corpus; lexical features; specialised corpus; language description; word lists analysis

INTRODUCTION

The tenet of a language description is in its words. Therefore, in the teaching and learning of English for Specific Purposes (ESP), the concern of many instructors is to ensure that the learners are exposed to the language that is specific to their disciplines. This involves the word and its lexical units, such as collocations and colligations.

Efforts have been made to explore the features of specialised languages to serve the language needs of learners from different domains. Corpus has been one of the tools to describe the features of a specialised language empirically (Lu et al. 2017, Peña & Peña 2015, Sadeghi & Nobakht 2014, Kashiha & Heng 2014, Peters & Fernández 2013, Kanoksilapatham 2013). Most corpus-based language investigations capitalise on word list analyses; among the most commonly employed are frequency, keyword, and key-keyword word lists (Partington & Marchi 2018, Rizzo & Pérez 2015, Lee 2014, Goh 2011). The frequency word list highlights the most frequent words in a corpus. The keyword list provides high occurrence words relative to the whole individual corpus; the words are said as to be specific to or representative of the target corpus, relative to another general corpus. The key-keywords list presents the most frequent keywords in a target corpus (Scott 1997).

Analyses from word lists have been used to extract distinguishing words of many specific-domain languages, such as Nelson (2000) for Business English, Fuentes and Fuentes

(2002) for Business and Computer Science English, Mudraya (2006) for Engineering English, and Lei and Liu (2016) for Medical English. The differences discerned from the word lists analyses allow observations of the specialised language constructions from word levels. Though the three word lists have been used in many corpus-based language studies, comparisons across these three types of word lists in characterising a specialised language has not been made to identify any salient information each word list can reveal about the target language. The discussions were more on how the employed word list in their studies characterise the target corpus, rather than looking into the different lexical information revealed by comparing the three word lists of the target language.

This paper demonstrates the extent to which each of these word lists can possibly be utilised in identifying and providing useful explanations in describing a specialised language. Therefore, the purpose of this paper is to investigate the lexical information that can be revealed about a specialised language from the analyses of the three types of word lists: frequency, keyword, and key-keyword. The observations from the analyses help to determine the different (or similar) lexical features derived from the three word lists. The knowledge on different lexical information that can be retrieved from each word list is helpful to inform language researchers the most appropriate word list to be employed for a specialised language investigation.

The paper is organised as follows. The next section provides the fundamental framework for extracting significant specialised words from the word lists for lexical profiling. The current study is described in the following section. The methodology section presents the corpora and word lists used for the study. Then, the results section presents the findings from the frequency, keyword and key-keyword word lists generated from the specialised corpus. Finally, the discussion section provides the comparison of the three word lists in providing the lexical information about the specialised language.

WORD LISTS FOR LEXICAL PROFILING: EXTRACTING SIGNIFICANT SPECIALISED WORDS

In a language investigation, word lists are generated as the starting point to determine the lexical units for further linguistic scrutiny (Noorli & Imran Ho 2015). Therefore, prior to the analyses of the word lists in this study, this section explains the fundamental principles in extracting significant words for the investigation of the specialised language.

The basic framework for extracting significant words from the specialised corpus in this study is adapted from Paquot (2005), who proposes a four-layered sieve to extract English for Academic Purposes (EAP) words. This framework provides the fundamentals of adopting the frequency, keyword and key-keyword lists for a language investigation.

The sieve consists of a series of quantitative filters: keyness, frequency, range and evenness of distribution (Figure 1). Though in her work Paquot suggested the use of lemma, this study resorts to the use of word forms because the researcher intends to identify all the significant word forms used in the specialised corpus. Sinclair asserts that "... anyone studying a text is likely to need to know how often each different word-form occurs in it" (1991, p. 30).

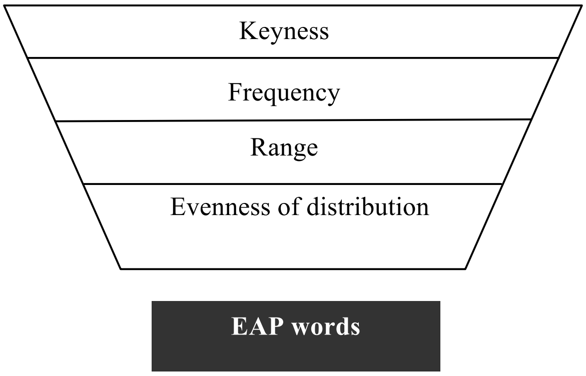

FIGURE 1. The four-layered sieve to extract EAP words

There have been arguments pertaining to the use of word forms versus lemmas for analysis in a language description (Cech & Macutek 2009). The head words in dictionary entries are lemmas. However, corpus data reveal that different word forms occur in different text types. Therefore, they have different collocational and colligational properties; this also means that they have different meanings (Scott 1997, Stubbs 1998). In a 200-million-word corpus, Stubbs (2001) found that different forms of *seek*, such as *seek*, *seeks*, *seeking* and *sought* have different collocates. A similar feature is also found in the collocation *seek-asylum*, which occurs in various forms: *asylum seekers*, *seeking asylum*, etc. Therefore, "... the unit of use and of meaning may be smaller than the lemma" (Stubbs 1998).

Stubbs (2009) further argues that because different lemmas have different frequency of occurrences, the use of lemmas as a linguistic unit is questionable. Gardner (2007) adds that the use of lemma for language investigation may dismiss crucial information, because English lexical units include multi-word items, such as prefabs and fixed phrases. Hence, the validity of word count and vocabulary study is, again, unreliable. Besides, the study conducted by Paquot involved a larger set of target corpus, with the Micro-Concord corpus collection B at 1,000,000 words of published academic prose. Hence, lemmatised items were preferable and manageable.

Elaborations on the Paquot's quantitative filters are provided in the following subsections.

KEYNESS

Paquot's (2005) first layer of EAP words extraction operates based on the concept of keyness – the high occurrence of words in a corpus in comparison with a reference corpus; this suggests the employment of the keyword list for a language investigation. For the keyword list analysis in this study, the specialised corpus is compared with a reference corpus to extract the specific words used in the specific domain (Figure 2). This is to identify the domain-specific words of the target corpus. The keyword list produced from this procedure provides the highly significant words in the specialised corpus (Goh 2011).

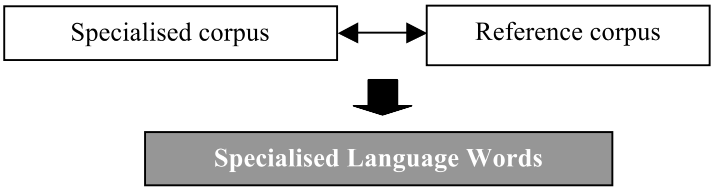

FIGURE 2. The comparison of keyword lists

The analysis employs the Log Likelihood test as the statistical measurement with the significance value set at 0.0000001.

FREQUENCY, RANGE AND EVENNESS OF DISTRIBUTION

In Paquot's study, general academic words were selected from keywords with frequencies that were equal to or higher than 30 occurrences from a 1,000,000-word corpus; this suggests the adoption of the key-keyword list for the language investigation. To distinguish words which occur frequently in most academic texts from others that were restricted to a specific discipline, the range criterion was considered, i.e. words appearing in all academic disciplines were retained as EAP vocabulary.

A similar procedure is adopted in this study with several adjustments (Figure 3). Because the study involves an investigation of only one specific discipline (Engineering), the interpretation of the EAP words for this study is the specialised language words. If the calculation in selecting the EAP words by Paquot (2005) is adopted, the specialised language words should be from the keywords with frequency of equal to or higher than 20 occurrences from 677,989 words. Since there are only two genres included in the specialised language of this study, the range criterion can be easily determined.

Meanwhile, the evenness of distribution, is reflected in the plot displays retrieved by the Wordsmith program serves as a supplementary means to provide the visual support for the words examination.

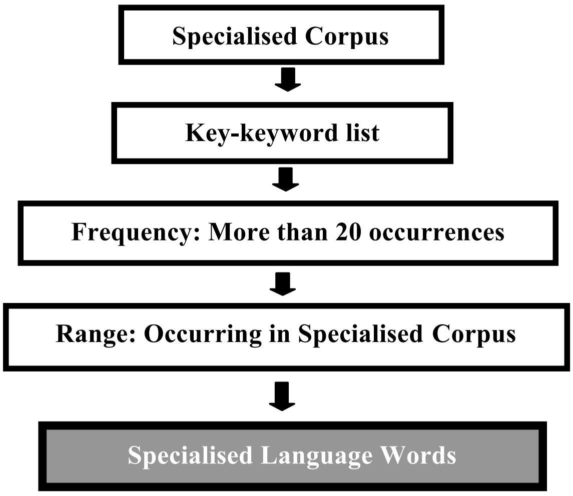
Specialised Corpus
Key-keyword list
Frequency: More than 20 occurrences
Range: Occurring in Specialised Corpus
Specialised Language Words

FIGURE 3. The application of frequency and range criteria

THIS STUDY

This study attempts to examine the profiling of a specialised language (Engineering English) depicted by the three word lists – frequency, keyword and key-keyword. Additionally, the analyses are carried out to determine any useful information that can be observed from each word list type in describing the specialised language. Hence, comparisons are carried out by looking at the distributions of the vocabulary components in the three word lists, in terms of:

i) function words and content words, and
ii) General Service List (GSL), Academic Word list (AWL), and Others

Distributions of these vocabulary components have been carried out in many corpus-based investigations which attempt to describe specialised languages (Noorli & Imran-Ho 2013, Gilmore & Millar 2018).

FUNCTION WORDS AND CONTENT WORDS

Though many studies excluded function words in describing a language, this one attempts to investigate function words used in this specific domain to discover possible features that significantly distinguish it from General English. Flowerdew (1997) asserts that function words are unique, because many of the members display a quality that joins grammar and lexis, such as the word *from*, which has 26 definitions in the COBUILD dictionary. Empirical observations of the function words also may lead to significant findings about rhetorical functions in ESP texts.

Function words have been variously defined as words which:

a. carry more structural information than semantic information,
b. are subject to more severe phonological modifications,
c. do not (are less likely to) bear pitch accents
d. are members of closed, (or nearly closed, e.g. numerals) classes.
e. are not nouns, main verbs, adjectives or some kinds of adverbs

Hence, function words comprise pronouns, prepositions, articles, conjunctions, and auxiliary verbs. Basically, they provide the "... cement that holds the content words together" (Chung & Pennebaker 2007, p. 347).

The content words are categorised according to General Service List (GSL), Academic Word List **(**AWL) and *Others*. *Others* include the technical, sub-technical and non-technical words. Non-technical words are general words which are not included in either GSL or AWL, such as *abrupt*, *accomplish* and *advantageou*s. Proper nouns, such as names of person, place and concepts are also classified under this category. This also implies that these non-technical words are infrequent words in general (GSL) and other academic (AWL) texts (Noorli & Imran Ho 2013).

GENERAL SERVICE LIST (GSL)

This list contains 2,000 words that are regarded as providing 'general service' to English learners. This list was published by Michael West in 1953. The selection of the words was based on written English; they are said to be the most frequent English words. Some ESL practitioners may regard this list as an essential list for English learners because a learner who can learn all the words in the list would be able to understand about 90-95% of colloquial speech and 80-85% of general written texts. Despite some sceptical observations made by some researchers on several issues related to this list, such as its adequacy and relevancy (because this list was issued in 1950s), many investigations into language description are still adopting this list especially for vocabulary profiling.

Because this list is out of print, there have been many versions with some improvements made to the list available. In this study, the version provided by Nation (Heatley et al. 2002) is employed.

ACADEMIC WORD LIST (AWL)

An academic word list developed by Coxhead (2000) was used for this study. This list stemmed from the need to prepare learners for academic study. Based on the principles of corpus linguistics, words which display commonness, with high frequency, in characterising academic activities such as research, analysis and evaluation across a wide range of academic sources were identified as academic words (Granger & Paquot 2009). These academic words were found infrequent in non-academic texts.

The list consists of 570 word families from a 3,500,000-word corpus of academic texts. These words are not included in West's 1953 first 2,000 words of GSL. The AWL has been widely used for language teaching, testing and material development.

The underlying principle for selecting this list is the nature of the words, which reflects the academic activities. Because this study aims to identify the characteristics of the Engineering English, the use of this list can shed some light particularly in characterising the corpus from the perspective of academic texts in terms of word category.

RESEARCH QUESTIONS

The research questions set for this study are:

a) How do the words in the frequency, keyword and key-keyword lists differ or similar?
b) How do the word lists differ, or similar, in terms of the distribution of:
   i) content and function words?
   ii) GSL, AWL and *Others* word categories?

METHOD

CORPORA FOR THE STUDY

Two corpora are used in this study – the Engineering English Corpus and British National Corpus (BNC)

The Engineering English Corpus, henceforth referred as E^2C, acts as the specialised corpus for the study. It comprises 102 texts with 677,993 of running words. This corpus was constructed from two academic Engineering sources – reference books and journal articles.

There were 34 chapters, with 425,854 running words, from the Engineering reference books. For manageability, the researcher selected only two textbooks - *Electronic Devices and Circuit Theory* (*8th edition*) and *Electronic Circuit Analysis* (*2nd edition*).

In addition, 68 journal articles, with 252,139 running words, were selected from the online databases. The search for articles from these databases was conducted by keying-in the key words from the chapter titles (of the two reference books) in the advance search column. The articles which appeared on the top list of the search results were given the priority for selection. Table 1 provides the composition of E^2C.

TABLE 1. The composition of E^2C

Sources	No. of texts	Running Words
Reference Books	34	425,854
Journal Articles	68	252,139
Total	102	677,993

The second corpus, BNC, is a reference corpus for the study. The comparison between the specialised corpus and the reference corpus aims to obtain the statistical information of the words from the Engineering English Corpus, thus, proving whether the identified words are specific to Engineering English (Meyer 2002). As such, BNC is regarded as a general English corpus in this study.

BNC frequency word list was used to provide comparison with the E^{2}C frequency word list. The reference corpus is also used when generating the keyword list of E^{2}C.

THE WORD LISTS: FREQUENCY, KEYWORD AND KEY-KEY-WORDS

The *Wordsmith* software provided the point of departure for the whole investigation. The three types of word lists generated with *Wordsmith* for this study were frequency, keyword and key-keyword. The frequency word list is generated from the word list program, and the keyword list from the keyword program. However, the key-keyword list is retrieved from the keyword database. The key-keywords are the most frequent keywords in a corpus or any set of files. Therefore, key-keywords are basically the most typical keywords in a corpus (Scott 1997).

OTHER WORD LISTS: BROWN FUNCTION WORDS, GSL AND AWL

The Brown Function Words was employed in the study. A list of 216 items or functions words were identified from the Brown corpus, and it was retrieved online from http://web.simmons.edu/~veilleux/ fw_project/ bcfw_list.htm. These function words constitute the most frequently occurring words in any texts. This list was used for the analyses of the distribution of function words in the E^{2}C. Other word lists were the word categories, GSL and AWL.

RESULTS AND DISCUSSION

FREQUENCY WORD LIST OF E^{2}C

Table 2 displays the top 30 words of E^{2}C and BNC. E^{2}C shows its characteristics as expected. The top nine words are function words, which cover nearly 28% of the corpus: *the*, *of*, *is*, *a*, *and*, *in*, *to*, *for* and *that*. The first content word *voltage* ranks as the 10th most frequent word, before the rest of the content words *current*, *circuit* and *output* appear in the top 20 words. More content words can be found after the top 30 words, such as *figure*, *gain*, *signal*, *resistance*, *source*, *power* etc. Therefore, like other corpora, function words still dominate the top frequent words in E^{2}C - to be exact 23 out of top 30 words (approximately 36% out of 41% of text coverage). A look at the content words in the top 100 of E^{2}C shows that the words are predominantly from the technical and/or sub-technical vocabulary. Apart from those which have been mentioned, it appears that most of the content words are nouns, such as *frequency, load, diode, device, emitter, value, level, network etc.* Words like *v*, *b* and *n* are used as symbols for devices, types or concepts such as *n-channel*, *amplifier b*, and *volts* (a measurement unit). It is found that the text coverage of the top 100 words from E^{2}C and BNC is about 54% and 46% respectively; the specialised corpus has more text coverage than BNC.

TABLE 2. E^{2}C and BNC frequency list (top 30)

		E^{2}C				BNC		
N	Word	Freq.	%	Cum. %	Word	Freq.	%	Cum %
1	THE	57,617	9.58	9.58	THE	6,055,105	6.09	6.09
2	OF	20,558	3.42	13.00	OF	3,049,564	3.07	9.15
3	IS	16,286	2.71	15.70	AND	2,624,341	2.64	11.79
4	A	15,432	2.57	18.27	TO	2,599,505	2.61	14.41
5	AND	15,378	2.56	20.83	A	2,181,592	2.19	16.60
6	IN	14,655	2.44	23.26	IN	1,946,021	1.96	18.56
7	TO	13,409	2.23	25.49	THAT	1,604,421	1.06	18.56
8	FOR	7,291	1.21	26.71	IS	1,052,259	0.98	19.61
9	THAT	6,233	1.04	27.74	IT	974,293	0.93	20.59

10	VOLTAGE	6,051	1.01	28.75	FOR	922,687	0.89	21.52
11	BE	5,708	0.95	29.70	WAS	880,848	0.87	22.41
12	AS	5,625	0.94	30.63	I	863,917	0.74	23.27
13	ARE	4,558	0.76	31.39	ON	732,523	0.74	24.01
14	CURRENT	4,347	0.72	32.11	WITH	731,319	0.66	24.75
15	WITH	4,320	0.72	32.83	AS	659,997	0.66	25.41
16	CIRCUIT	4,064	0.68	33.51	BE	655,259	0.66	26.07
17	THIS	3,942	0.66	34.16	HE	651,535	0.60	26.72
18	OUTPUT	3,753	0.62	34.79	YOU	593,609	0.59	27.32
19	BY	3,645	0.61	35.39	AT	588,503	0.53	27.91
20	AT	3,300	0.55	35.94	BY	524,075	0.52	28.44
21	AN	3,293	0.55	36.49	ARE	513,444	0.46	28.96
22	CAN	3,235	0.54	37.03	THIS	458,368	0.46	29.42
23	WE	3,033	0.50	37.53	HAVE	454,419	0.45	29.87
24	WILL	2,949	0.49	38.02	BUT	448,684	0.45	30.32
25	ON	2,899	0.48	38.50	NOT	446,783	0.43	30.77
26	INPUT	2,822	0.47	38.97	FROM	431,075	0.43	31.21
27	OR	2,758	0.46	39.43	HAD	425,987	0.42	31.63
28	FROM	2,577	0.43	39.86	HIS	413,144	0.41	32.05
29	FIG	2,433	0.40	40.26	THEY	410,294	0.38	32.46
30	TRANSISTOR	2,391	0.40	40.66	OR	376,289	0.37	32.84

Further comparison with the top 100 frequent words from BNC reveals different words between the Engineering English Corpus and general English. The words in BNC are more general, and the function words reign supreme even within the top 100 frequent words; at least the first 50 most frequent words are function words.

TABLE 3. E^{2}C top 30 function words

THE	AS	ON
OF	ARE	OR
IS	WITH	FROM
A	THIS	WHICH
AND	BY	IT
IN	AT	IF
TO	AN	HAVE
FOR	CAN	THEN
THAT	WE	THAN
BE	WILL	ONE

There are 213 function words found in E^{2}C. Figure 4 displays the distribution of function to content words in E^{2}C in percentage. These function words cover almost 44% of the corpus. Table 3 presents the first 30 function words identified. The occurrence of these words needs to be examined with caution because some of these words do not behave as function words all the time, for example *is*, which can also be a verb. However, in this study, all the identified functions words are not edited, and all are treated as words that match the function words from the Brown Corpus Function Word list.

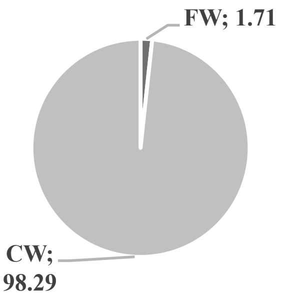

FIGURE 4. The distribution of function to content words in E^{2}C (%)

Several types of function words can be identified from the list including articles, prepositions, conjunctions, modals, auxiliary verbs and pronouns.

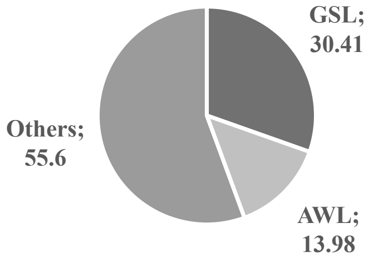

FIGURE 5. The distribution of GSL, AWL and *Others* in E^2C (%)

Next, the content words are categorised according to GSL, AWL and *Others*. Figure 5 shows the distribution of the categories in percentages. The *Others* category has the highest number of word types (around 56%), suggesting that E^2C is indeed a specialised language. A closer examination of the word categories reveals that there are overlapping words between the *Others* category and GSL, and the *Others* category and AWL. In other words, there are some words occurring in GSL and AWL which carry some degree of technical senses (sub-technical words) in E^2C, such as *base*, *bias*, *channel*, *collectors* and *field*. However, in this study, there is no attempt to edit the lists to identify and extract all those words. The intention is to determine the proportion of words in E^2C that are listed in GSL and AWL. Of course, if the extraction and classification are to be done, it results in different proportions in the overall distributions. A larger fraction of *Others* and smaller proportions of GSL and AWL may prevail. Because the scope of this research does not include the investigation on the classification of technical, sub-technical and non-technical words in the corpus, the distributions in Figure 5 is taken as a preliminary insight into the lexical profiling of the target corpus. Furthermore, Fraser (2005) claims that sub-technical words (general cryptotechnical and lay-technical words) occur with higher frequency than technical words which are specific to the field. Mudraya (2006) adds that the non-technical sense of a sub-technical words are used more frequently than its technical sense. Because these words have higher frequency of occurrence, they stand out in the frequency word lists.

KEYWORD LIST OF E^2C

This section examines the significant words occurring in E^2C. The specialised corpus was run against the reference corpus for this study, the BNC, which is also taken as the General English. These are the words that characterise E^2C through their high occurrence relative to the whole individual corpus.

Table 4 lists the top 30 keywords in E^2C, ordered according to their keyness. The threshold set for the computation of the data is the minimum significance value at 0.0000001 (log likelihood) and minimum frequency at 1 (to retrieve all the keywords). The log-likelihood score highlights the significant value of a word, which informs the distinctiveness of the word when its usage in the target corpus is compared with its usage in the reference corpus.

There are 3,100 key words identified, with the text coverage total of 80.26%. Positive keywords form a total of 2,196 items, which constitute about 64% of text coverage, and the remaining 903 negative keywords make up another 16% of text coverage. Positive keywords occur more often than would be expected by chance in E^2C in comparison with BNC; conversely, negative keywords occur less often in E^2C than would be expected by chance in comparison with BNC. Table 4 lists both the positive and negative keywords. The negative keywords are reordered from the most negative keyword.

TABLE 4. E^{2}C positive and negative keyword list (top 30)

N	Positive Keywords			Negative Keywords		
	Keyword	%	Keyness	Keyword	%	Keyness
1	VOLTAGE	1.01	54921.54	HE	0.01	-7713.71
2	CIRCUIT	0.68	31607.73	I	0.07	-7265.69
3	OUTPUT	0.62	24422.70	WAS	0.15	-6640.99
4	CURRENT	0.72	23327.13	HIS	0.00	-5444.95
5	TRANSISTOR	0.40	22160.54	YOU	0.07	-5325.01
6	INPUT	0.47	19273.35	HAD	0.01	-4907.54
7	SIGNAL	0.36	14492.23	SHE	0.00	-4279.00
8	GAIN	0.37	13203.43	IT	0.33	-4180.60
9	FIG	0.40	13150.82	HER	0.00	-4050.05
10	AMPLIFIER	0.25	12966.52	THEY	0.07	-2908.73
11	DIODE	0.23	12900.96	SAID	0.00	-2549.92
12	EMITTER	0.20	11400.91	BUT	0.14	-2378.07
13	RESISTANCE	0.28	10255.42	WHO	0.01	-2367.02
14	CIRCUITS	0.20	9713.64	HIM	0.00	-2036.62
15	IS	2.71	9705.26	WHAT	0.03	-2022.44
16	FREQUENCY	0.24	9186.69	MY	0.01	-1724.61
17	TRANSISTORS	0.16	8928.36	WERE	0.09	-1659.60
18	LOAD	0.24	8884.35	ME	0.00	-1631.17
19	FIGURE	0.37	8733.42	THEIR	0.06	-1562.20
20	DEVICE	0.22	8193.59	THEM	0.02	-1538.01
21	BIAS	0.18	7471.03	IT'S	0.01	-1472.05
22	SOURCE	0.26	6875.59	OUT	0.04	-1381.89
23	DC	0.18	6854.01	DO	0.03	-1314.29
24	COLLECTOR	0.16	6692.32	PEOPLE	0.01	-1306.26
25	SHOWN	0.29	6487.58	YOUR	0.01	-1305.83
26	RESISTOR	0.11	6007.45	BEEN	0.08	-1291.18
27	OP	0.13	5829.43	TO	2.23	-1290.25
28	THE	9.58	5658.11	UP	0.05	-1244.28
29	CAPACITOR	0.11	5628.93	THERE	0.11	-1180.41
30	CONFIGURATION	0.13	5469.74	KNOW	0.01	-1168.52

As can be seen, the positive keyword list provides more specific and technical words occurring in E^{2}C. Generally, this keyword list reveals that nouns (e.g. *voltage*, *circuit*, *outputs*, *capacitance* and *device*) dominate the top 30 words. There are also some abbreviations included in this list, such as *Fig*, *DC*, *OP*, *AC*, *MOSFET* and *CMOS*. The occurrence of these abbreviations as keywords in E^{2}C suggests that one of the main characteristics of this specific domain is the use of such abbreviations to represent concepts.

Nelson (2000) notes that it is possible to describe the language in a specific domain by investigating 'what is *not* found there'. It can be achieved by using the negative keywords. The E^{2}C negative keyword list mostly contains function words. In fact, its top 10 words are from this word category: *he*, *I*, *was*, *his*, *you*, *had*, *she*, *it*, *her* and *they*. The rest of the words include more general words, such as *people*, *years*, *year* and *out*. The occurrence of some de-lexicalised verbs (verbs which have low lexical content and their meanings in contexts are conditioned by other co-existing words), such as *know* and *got*, further distinguishes E^{2}C as a language of a specific domain. The contrast difference between the positive and negative keywords underlines the 'specialised' characteristic of E^{2}C.

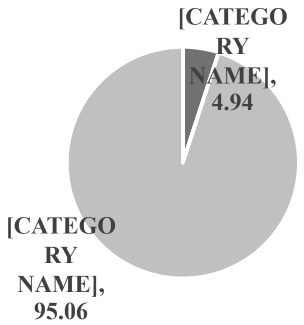

FIGURE 6. The distribution of function and content words in E^2C keyword list (%)

Unlike the frequency word list, only 1 function word makes it to the top 30 positive keyword list, namely *is*. However, in the whole keyword list, there are 154 function words identified. Figure 6 displays the distribution of the function words to content words in E^2C keyword list. It reveals that the proportion of function words in the keyword list is higher (5%) than in the frequency word list (2%). Out of this 5%, there are only 31 positive key function words. The positive key function words still include prepositions, articles, modals, conjunctions, pronouns, and auxiliary verb. The identification of these function words as keywords signifies that they are worthy of a closer examination to see how they contribute to the characterisation of this specific domain. The list of 31 positive key function words and 31 negative key function words, sequenced from the lowest keyness, is shown in Table 5. The table reveals that pronouns (*I*, *his*, *you*, *she*, *it*, *her*, *they*, *him*, *my*, *me*) are unlikely to characterise E^2C. The occurrence of *is* and *are* in the positive keyword list, and *was* and *were* in the negative keyword list suggests that past tenses are less likely to be a feature of E^2C. The occurrence of *had* as a negative keyword supports this notion as well.

TABLE 5. E^2C positive and negative key-function-words

Positive Keywords			Negative Keywords		
Keyword	%	Keyness	Keyword	%	Keyness
IS	2.71	9705.26	I	0.068	-7265.69
THE	9.58	5658.11	WAS	0.154	-6640.99
CAN	0.54	1590.17	HIS	0.003	-5444.95
SINCE	0.17	847.39	YOU	0.072	-5325.01
THEREFORE	0.11	789.87	HAD	0.015	-4907.54
VERSUS	0.03	739.15	SHE	0.003	-4279.00
WILL	0.49	695.81	IT	0.328	-4180.60
ARE	0.76	535.18	HER	0.002	-4050.05
ACROSS	0.09	514.67	THEY	0.066	-2908.73
WE	0.50	381.26	BUT	0.135	-2378.07
AN	0.55	341.98	WHO	0.005	-2367.02
EACH	0.14	337.44	HIM	0.001	-2036.62
BE	0.95	297.15	WHAT	0.028	-2022.44
HOWEVER	0.13	289.77	MY	0.006	-1724.61
BETWEEN	0.18	246.42	WERE	0.092	-1659.60
AS	0.94	246.36	ME	0.003	-1631.17
WHEREAS	0.03	237.97	THEIR	0.065	-1562.20
FOR	1.21	224.20	THEM	0.019	-1538.01
THIS	0.66	188.68	IT'S	0.005	-1472.05
TOWARD	0.01	167.18	OUT	0.041	-1381.89
MUST	0.13	133.08	DO	0.033	-1314.29
OPPOSITE	0.02	129.02	YOUR	0.013	-1305.83

ABOVE	0.06	126.23	BEEN	0.085	-1291.18
THAN	0.22	109.55	TO	2.229	-1290.25
IN	2.44	106.56	UP	0.055	-1244.28
BELOW	0.03	91.69	THERE	0.110	-1180.41
THEN	0.23	84.71	HAVE	0.234	-1158.54
THROUGH	0.12	52.38	LIKE	0.028	-1098.05
OFF	0.11	50.30	ON	0.482	-1087.77
MINUS	0.01	45.07	DON'T	0.004	-1065.65
EITHER	0.05	34.31	NO	0.080	-1057.48

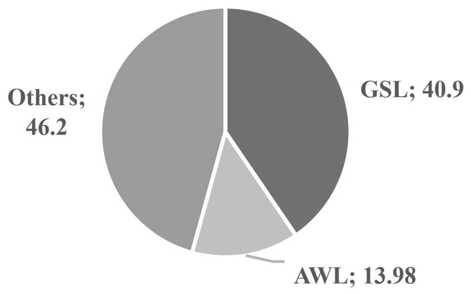

FIGURE 7. The distribution of GSL, AWL and *Others* in E^2C keyword list (%)

The content words are also re-examined to observe any dissimilarity present in the distribution of GSL, AWL and *Others* categories in comparison to the frequency word list. The distribution of the categories is illustrated in Figure 7. A cross-reference with Figure 5 reveals that the keyness notion has resulted in some degree of adjustment to the distribution of categories in the corpus; it displays almost a balanced use of GSL (41%) and *Others* (46%). There is a slight reduction in the use of AWL, by around 1%. This means that the keyword list provides a different set of words and categories in describing the lexical profile of a corpus. The words are more specific, thus, there is a lesser number to be analysed, and most importantly, their occurrences are significant in the corpus.

KEY-KEYWORD LIST OF E^2C

The keyword list allows the formation of a keyword database, which reveals the key-keywords of a corpus. The key-keyword list in turn enables the identification of a word range – how many texts in the corpus does the word occur in. The more texts it is 'key' in, the more 'key-key' it is. The selection of words for analysis in a language normally is based on the *frequency* and *range* criteria (Utimaya & Chujo 2007, Paquot 2005). Apart from that, the keyword database also provides 'associates' of a key-keyword – other keywords that occur in the same texts as the key-keyword is. Once again, this information is helpful for word analyses.

This list provides keywords which appear in 3 texts and more in E^2C. There are 916 key-keywords in the list. The first 30 key-keywords (Table 6) still exhibits the dominance of nouns in E^2C. Nevertheless, more verbs are listed in its top 100 words, such as *shown*, *shows*, *determine*, *using*, *applied*, and *connected*. Some symbols and abbreviations are also still making up the feature of E^2C in this list: *fig*., *DC*, *AC*, *B*, *N*, *BJT*, *V*, *IC* and *MOSFET*. Several adjectives are also making their way to the list: *low*, *high*, *negative*, *maximum*, *constant*, and *linear*. It seems that more word classes are included in the top 100 list. This finding suggests that the key-keyword list offers more classes of words for a specialised language investigation.

TABLE 6. Key-keyword lists of E^2C (top 30)

N	KW	Texts	%	Overall Freq.	N	KW	Texts	%	Overall Freq.
1	VOLTAGE	71	69	6026	51	EXAMPLE	30	29	1026
2	CIRCUIT	70	68	4046	52	N	30	29	806
3	IS	64	62	14926	53	TEMPERATURE	30	29	697
4	OUTPUT	62	60	3696	54	AMPLIFIERS	29	28	272
5	CURRENT	61	59	4290	55	B	29	28	882
6	TRANSISTOR	58	56	2378	56	BIASED	29	28	679
7	CIRCUITS	55	53	1172	57	SOLUTION	29	28	524
8	SHOWN	53	51	1642	58	DETERMINE	27	26	545
9	SIGNAL	50	49	2122	59	GATE	27	26	905
10	INPUT	48	47	2748	60	JUNCTION	27	26	528
11	AMPLIFIER	47	46	1457	61	MAGNITUDE	27	26	375
12	TRANSISTORS	47	46	949	62	MAXIMUM	27	26	510
13	DIODE	46	45	1359	63	NEGATIVE	27	26	441
14	FIG	46	45	2407	64	TERMINAL	27	26	413
15	GAIN	46	45	2183	65	EQUATIONS	26	25	352
16	BIAS	45	44	1046	66	RESULTING	26	25	458
17	DC	45	44	1032	67	ZERO	26	25	522
18	PARAMETERS	45	44	727	68	APPLIED	25	24	649
19	RESISTANCE	45	44	1649	69	EQUIVALENT	25	24	731
20	EMITTER	44	43	1173	70	LINEAR	25	24	279
21	FIGURE	44	43	2071	71	RATIO	25	24	241
22	VOLTAGES	44	43	524	72	SIMULATION	25	24	218
23	DEVICE	43	42	1281	73	V	25	24	1027
24	RESISTOR	43	42	652	74	CAPACITANCE	24	23	395
25	SOURCE	42	41	1394	75	FUNCTION	24	23	573
26	THE	42	41	44190	76	RESULTS	24	23	545
27	FREQUENCY	41	40	1402	77	SIGNALS	24	23	268
28	LOAD	40	39	1419	78	APPLICATIONS	23	22	411
29	VALUE	39	38	940	79	CONSTANT	23	22	385
30	CAPACITOR	38	37	638	80	DRAIN	23	22	548

In this key-keyword list, the use of function words covers 3% (28 words) of the total word types (Figure 8). The function words included in this list is as shown in Table 7. The list comprises some modals, *can*, *may*, *must* and *will*, apart from other auxiliary verbs, *are*, *be*, and *is* (which need to be distinguished from the real verb), and prepositions, *above*, *across*, *for*, *of*, and *off*. Markers such as *however*, *or*, *since*, *then*, and *whereas* are also listed in this key-keyword list. The inclusion of more types of function words in this list suggests the suitability of the key-keyword list for further linguistic analyses of a specialised language.

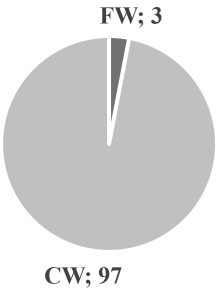

FIGURE 8. The distribution of function and content words in E^2C key-keyword list (%)

TABLE 7. Key-key-function-words of E^2C

A	IN	THESE
ABOVE	IS	THIS
ACROSS	MAY	VERSUS

AN	MUST	WE
ARE	OF	WHEREAS
BE	OFF	WILL
CAN	OR	
EACH	SINCE	
FOR	THE	
HOWEVER	THEN	

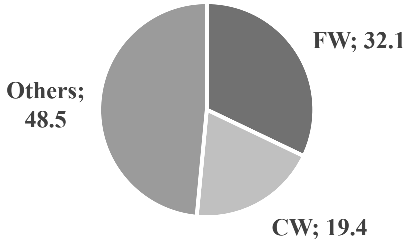

FIGURE 9. The distribution of GSL, AWL, and *Others* in E^{2}C key-keyword list (%)

Though the proportions of GSL, AWL and *Others* categories in the E^{2}C key-keyword list distribution (Figure 9) are different from the keyword list, *Others* still has the largest number of words in the list, followed by GSL and AWL. AWL, nonetheless, has a higher proportion in the key-keyword list than in the keyword list; it is caused by the threshold set at more than 2 texts of occurrence, which subsequently results in the smaller proportion of *Others*. Similarly, the distribution of these word categories proves that the key-keyword list provides an appropriate set of words to receive priority for closer investigation on the lexical behaviour of E^{2}C.

DISCUSSION

The frequency word lists highlight the specific features of E^{2}C as opposed to GE, with lesser function words occurring in its high frequency words. The higher text coverage of the top 30 words, that is 41% in E^{2}C and only 33% in BNC further proves this. The specific topics or areas in E^{2}C allow the use of more specific and less varied words across the corpus.

The keyword list provides different, but more detailed features of the specialised corpus. It exposes more specific and technical words in all the corpora. Though nouns still dominate the top 30 positive keywords, abbreviations and symbols appear to claim a place in describing the feature of the specific domain. However, function words, particularly pronouns, dominate the negative keyword lists; hence, it suggests that function words do not make up the characteristics of E^{2}C. The occurrence of past tenses in the negative keyword also rules out the tense as a feature of the specific domain.

The key-keyword list offers more varied members in the top list. The list comprises lesser number of words, but still demonstrates the dominance of nouns, with inclusions of verbs and adjectives, while retaining a few function words and more abbreviations and symbols. In other words, the lists provide a good range of words, which entails the priority for analyses in describing the characteristics of E^{2}C. In fact, Paquot's (2005) first three criteria of EAP word selection: keyness, frequency and range, are performed in the key-keyword list procedure.

The discussion of word lists thus far has proven that the different word lists analyses result in different range of words to be discovered.

FUNCTION WORDS

TABLE 8. Distribution of function words to content words in E^{2}C word lists

	Frequency Word List	Keyword List	Key-Keyword List
Function Words (%)	1.71	4.94	3
Content Words (%)	98.29	95.06	97

Table 8 records the distribution of function to content words in all the word lists. Content words suggest the contents of the specialised language, while the function words suggest the style.

It shows that the frequency word list makes up the lowest distribution of function words, and the keyword list the highest. Despite the fact that the distribution in the keyword list includes the negative keywords, the proportion of positive key function words is bigger than the negative function words, and it is still the highest of all the lists. However, the key-keyword list includes significant function words which occur in more than 2 texts in the specialised corpus. Furthermore, unlike frequency and keyword lists, the higher ranked key-keywords cover a wide range of function words including prepositions, pronouns, conjunctions, modals and other auxiliary verbs. From the comparison, the distribution of the function words in the key-keyword list is reasonable, because it is slightly more than the frequency word list, and slightly lesser than the keyword list. All these qualities suggest that the key-keyword list makes a suitable list for a further investigation into the lexical behaviour of E^{2}C.

GSL, AWL AND *OTHERS*

It is mentioned earlier that some of the sub-technical words are listed in both GSL and AWL, and the fact that Granger and Paquot (2009) urge the need to be careful when using the GSL for ESP. Though this study thus far has classified the words into GSL, AWL and *Others*, the concern for the moment, however, is not on the identification and classification of the sub-technical words which involve the extractions of technical sense words from GSL and AWL. In other words, the aim is not to identify technical vocabulary. This study merely involves the classification of the words in the corpus according to their superficial quality to obtain the overview of the lexical profiles of the specialised corpus, such as how many of the words appear in the GSL and AWL, and how much they cover the corpus. The remaining words which do not fall into either category are classified as *Others*. This classification, though superficial, it provides some insights into the comparison of the general distribution of words in the corpus, which are quite practical for language instructors' immediate reference. Nation (2001b) proposes that the combination of the words in the GSL and AWL with words in the specific discipline should reach the critical 95% coverage threshold for reasonable reading comprehension. In addition, the identification and extraction of technical sense words from the GSL and AWL requires some degree of technical knowledge of the subject field. Further work on the identification and classification of technical vocabulary can be conducted following up this study.

Table 9 shows the comparison of the distribution of GSL, AWL and *Others* categories in the corpus.

TABLE 9. Distribution of GSL, AWL and *Others* in E^{2}C word lists

	Frequency Word List	Keyword List	Key-Keyword List
GSL (%)	30.41	40.9	32.1
AWL (%)	13.98	12.9	19.4
Others (%)	55.6	46.2	48.5

The general distributions of GSL, AWL and *Others* categories prove that this specialised corpus has different proportions of vocabulary types from general English. As proposed by Nation (2001a), high frequency words (GSL) constitute 80% of tokens in a text (corpus), academic words (AWL) make up 9%, technical and low frequency words contribute another 5% each. This great difference indeed entails a different approach not only in the study of the specialised language, but also in the teaching and learning of the language (Gavioli 2005).

Table 9 shows that the GSL, AWL and *Others* categories have a similar order of the proportions for E^2C in all the word lists: *Others,* GSL, AWL. The highest proportion of *Others* category reflects the specific features of the corpora because *Others* includes technical, sub-technical and non-technical words. However, a closer look at the words reveals that there are overlappings of categories with some of the words; some GSL and AWL words are found to be sub-technical. Hence, the proportion of *Others* category should be larger.

CONCLUSION

The observation from the three types of word lists analyses reveals a different range of words, and the analyses of the words reveal the distinct features of the specialised language at different levels. More word classes are revealed from frequency, keyword to key-keyword lists. Though the proportion of the vocabulary types – GSL, AWL and *Others*, is consistent in the three-word lists, the fact that *Ohers* is the highest word category in all the word lists emphasizes the specific feature of E^2C in comparison to General English.

This study also demonstrates the fact that out of the three word lists, the key-keyword list not only provides a wider word range, but also indicates the key-keyness of the word in a specialised corpus; the more texts a word is 'key' in, the more 'key-key' it is. Thus, it further substantiates the significance of the words in the specialised language. Besides, the list contains a manageable number of words; it promises a more effective selection of words to be studied in a specialised language investigation.

In conclusion, this study provides insights into the lexical information revealed in the three word lists. As such, it further verifies the usefulness of corpus-based investigations in providing empirical evidences for language description.

ACKNOWLEDGEMENTS

This paper is funded by a grant, GLuar/MPM/2017/PBPI-CTED/I00028, from Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

Cech, O. R. & Macutek, G. J. (2009). Word form and lemma syntactic dependency in czech: a comparative study. *Glottometrics.* 85-98.

Chung, C. & Pennebaker, J. (2007). The psychological functions of function words. In Fiedler, K. (Ed.). *Social Communication* (pp. 343-359). New York: Psychology Press**.**

Coxhead, A. (2000). The academic word list: a corpus-based word list for academic purposes. Paper presented at the 4th International Conference on Teaching and Language Corpora. Atlanta.

Flowerdew, L. J. (1997). Corpus linguistics: applications to ESP. Paper presented at Exploring Language 1997. Language Centre: HKUST.

Fraser, S. (2005). The lexical characteristics of specialized texts. Paper presented at JALT2004. Tokyo.

Fuentes, B. C. & Fuentes, A. C. (2002). A current corpus of technology language in Spain: English words that matter. *English for Specific Purposes World.* Retrieved February 20, 2011 from http://www.esp-world.info/Curado.htm.

Gardner, S. (2007). Integrating ethnographic, multidimensional, corpus linguistic and systemic functional approaches to genre description: an illustration through university history and engineering assignments. Paper presented at the 19th European Systemic Functional Linguistics Conference and Workshop.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins Publishing Company.

Gilmore, A. & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes. Vol 51*, 1-17.

Goh, G.Y. (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research. Vol 28* (1), 239-256.

Granger, S. & Paquot, M. (2009). In search of a General Academic vocabulary: a corpus-driven study. Paper presented at the International Conference 'Options and Practices of L.S.A.P practitioners*'*. Crete: University of Crete, February.

Heatley, A., Nation, I. S. P. & Coxhead, A. (2002). Range (Computer software). Retrieved July 25, 2011 from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Kanoksilapatham, B. (2013). Generic characterisation of civil engineering research article abstracts. *3L: The Southeast Asian Journal of English Language Studies. Vol 19*(3), 1-10.

Kashiha, H. & Heng, C. S. (2014). Discourse functions of formulaic sequences in academic speech across two disciplines. *GEMA Online Journal of Language Studies. Vol 14*(2), 15-27.

Lee, H. K. (2014). Phraseological patterns of English adjectives and nouns: with reference to the noun collocates of new, good, old and high in American English. *Linguistic Research. Vol 31*(3), 541-567.

Lei, L. & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes. Vol 22*, 42-53.

Lu, W., Lee, S.-M. & Jhang, S.E. (2017). Keyness in maritime institutional law texts. *Linguistic Research. Vol 34*(1), 51-76.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction.* Cambridge: CUP.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes. Vol 25*, 235-256.

Nation, I. S. P. (2001a). How many high frequency words are there in English? In Gill, M., Johnson, A.W., Koski, L. M., Sell, R. D. & Wårvik, B. (Eds.). *Language, Learning and Literature: Studies Presented to Håkan Ringbom English Department Publications 4* (pp. 167-181). Turku: Åbo Akademi University.

Nation, I. S. P. (2001b). *Learning Vocabulary in Another Language.* Cambridge: CUP.

Nelson, M. (2000). A corpus-based study of the lexis of Business English and Business English teaching materials. Unpublished PhD thesis, University of Manchester.

Noorli Khamis & Imran Ho-Abdullah. (2015). Exploring word associations in academic engineering texts. *3L: The Southeast Asian Journal of English Language Studies. Vol 21*(1), 117-131.

Noorli Khamis & Imran Ho-Abdullah. (2013). Word lists analysis: specialised language categories. *Pertanika Journal of Social Sciences & Humanities. Vol 21*(4), 1563-1581.

Paquot, M. (2005). Towards a productively-oriented academic word list. Paper presented at the *Corpora and ICT in Language Studies PALC 2005.* Frankfurt: University of Lodz.

Partington, A. & Marchi, A. (2018). Using corpora in discourse analysis. Paper presented at the *Corpora and Discourse International Conference.* Lancaster: Lancaster University.

Peña, G. A. & Peña, C.N. (2015). Extraction of candidate terms from a corpus of non-specialized, general language. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información. Vol 29*(67), 19-45.

Peters, P. & Fernández, T. (2013). The lexical needs of ESP students in a professional field. *English for Specific Purposes. Vol 32*, 236–247.

Rizzo, C. R. & Pérez, M. J. (2015). A key perspective on specialized lexis: keywords in telecommunication engineering for CLIL. *Procedia - Social and Behavioral Sciences. Vol 198*, 386-396.

Sadeghi, K. & Nobakht, A. (2014). The effect of linguistic context on efl vocabulary learning. *GEMA Online Journal of Language Studies. Vol 14*(3), 65-82.

Scott, M. (1997). PC analysis of key words -- and key key words. *System. Vol* 25(1), 1-13.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. (1998). A note on phraseological tendencies in the core vocabulary of English. The Free Library. Retrieved June 5, 2000 from http://www.thefreelibrary.com/A note on phraseological tendencies in the core vocabulary of English.-a093027799.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantic*s. Oxford: Blackwell.

Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics. Vol* 30(1), 115-137.

Utimaya, M. & Chujo, K. (2007). Linking word distribution to technical vocabulary. *Journal of the College of Industrial Technology. Vol 40*, 13-21.