Pertanika J. Soc. Sci. & Hum. 21 (4): 1563 - 1581 (2013)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Wordlists Analysis: Specialised Language Categories

Noorli Khamis[1]* and Imran Ho Abdullah[2]

[1]Centre of Languages and Human Development, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76109 Durian Tunggal, Melaka, Malaysia
[2]Pusat Pengajian Bahasa dan Linguistik (PPBL), Fakulti Sains Sosial Dan Kemanusiaan, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ABSTRACT

The study of lexical and grammatical patterns in the language that a learner must assimilate is important for pedagogical considerations. Therefore, the linguistic approach in ESP classrooms today is gaining its momentum. Additionally, the adoption of corpus-based language investigation has made the attempts even more accessible; many aspects of the specialised language can be described empirically and systematically. It has been discovered that word and structure frequencies of a specialised corpus are greatly different from a large corpus. They provide insights into the features of the specialised language. Hence, this paper demonstrates the different, yet useful information about a specialised language that can be discerned from the analysis of three types of wordlist; namely, frequency wordlist, keyword list and key-keyword list. The findings inform the features of the specialised language that need to be highlighted in an ESP classroom.

Keywords: Corpus, frequency wordlist, key-keyword list, keyword list, specialised language

INTRODUCTION

The study of ESP language features, in particular the lexical features, can facilitate the effective teaching of ESP. Hunston (2000) asserts that the area of ESP learning focus should be on words together with their companies. It is because a significant amount of language is constructed from sequences of more or less fixed forms of morphemes. These sequences finally form formulaic language distinctive to a particular language or register. If these formulaic constructions are highlighted in their contexts to the learners, learning can be reinforced. Noorzan (2005) further stresses that the study of lexical and grammatical patterns in the language that a learner must

ARTICLE INFO

Article history:
Received: 25 October 2011
Accepted: 28 August 2012

E-mail addresses:
noorli@utem.edu.my (Noorli Khamis),
imranho@ukm.my (Imran Ho Abdullah)
* Corresponding author

ISSN: 0128-7702

assimilate is important for pedagogical considerations.

Additionally, the adoption of corpus-based language investigation has made the attempts to discover the lexical features of a specialised language even more exciting; many aspects of the specialised language can be described empirically and systematically. Gavioli (2005) regards ESP and specialised corpora as one happy marriage. Specialised corpora are designed with the aim to represent a sub-language and to reflect the specific purpose of a research or teaching condition. The collection of texts may be from:

a. similar contents such as science, medicine, business or philosophy, **or**
b. from similar text-type/ genre such as research papers, letters or books, **or**
c. both; such as medical research articles or science lectures, **or**
d. texts from other types of specialised categories such as newspaper language or academic language

This study is an effort to look into the application of corpus work in describing a specialised language from the scope of vocabulary types or language categories – AWL, GSL and *Others*. The focus of the paper is on the analysis of three different types of wordlists, namely, frequency wordlist, keyword list and key-keyword list of a specialised corpus. This paper aims to demonstrate the different, yet useful information about a specialised language that can be discerned from the analysis of these wordlist types. The findings inform the features of the specialised language that need to be highlighted in ESP/EAP classrooms.

ESP VOCABULARY CATEGORIES

In vocabulary studies, Nation's (2001) classification of vocabulary types in a text has been cited by many researchers (Chung & Nation, 2003; Tsubaki, 2004; Martinez *et al*., 2009). The classification of the vocabulary types is summarised in Table 1.

However, for the description of science language, there is yet a clear cut classification of the vocabulary types. The confusion mainly involves defining words constituting academic words in the specialised language. Many question the accuracy of the AWL in the ESP pedagogic context because all subject-specific course types have their own set of lexical profiling (Granger & Paquot, 2009). According to Martin (1976), academic words are words which characterise academic activities such as the research procedure, analysis and evaluation. However, the computerised approach to quantify vocabulary of specialised texts has seen the classification of ESP vocabulary in the texts into technical, sub-technical (academic words), semi-technical and non-technical words (Menon & Mukundan, 2010; Fraser, 2007; Mudraya, 2006; Baker, 1988; Perez-Parades, 2003; Worthington & Nation, 1996).

Another issue regarding the distribution of the vocabulary categories is discussed by Mukundan and Aziz (2009) who caution material developers that high coverage of

TABLE 1
A summary of Nation's classification of vocabulary types

Types	Description
High frequency words	• words listed in the General Service Word List (GSL) • constitutes 80% of the running words in a text
Academic words	• words listed in the Academic Word List (AWL) • words that frequently appear in academic texts, but infrequently in non-academic texts • constitutes 9% of the running words in a text
Technical words	• include words common in a particular subject, but not others • constitutes 5% of the running words in a text • include words ranging from those not occurring in other subject areas to those with high frequency • these words carry specialised meaning • some may occur as high frequency or academic words
Low frequency words	• constitute the largest word group • constitutes 5% of the running words in a text • include words which do not fall under any types mentioned above.

(Source: Nation, 2001)

GSL or high frequency words in a corpus does not guarantee that learning can take place effectively. The distribution of these words needs to be taken into consideration too when analyzing the features of a corpus so that the right emphasis can be considered for pedagogical purposes.

RESEARCH QUESTIONS

The study aims to answer the following research questions:

a. What information can the analysis of different types of wordlists (frequency, keyword and key-keyword lists) from the same specialised corpus offer?

b. Do the different types of wordlists from the same specialised corpus affect the proportion of language categories (GSL, AWL and *Others*)?

METHODOLOGY

Corpora for the Study

There were two corpora used for this study; one which reflected the specialised corpus - Reference Books Corpus (RBC), and another reflected the general English, as well acted as the reference corpus - the British National Corpus (BNC). RBC was created by the researcher, while BNC was obtained online.

Reference Book Corpus (RBC)

The texts for the creation of the corpus for the study were identified from the Handbook of the Faculty of Electronics and Computer Engineering of one technical university located in Melaka. The handbook contains suggested textbooks for the students from all the programmes in the faculty. For manageability, the researcher selected only two suggested textbooks from a subject, which is a compulsory subject for all the first

year degree students of the faculty, regardless of their different programmes. In order to ensure that the books are the students' main references, they should be suggested as the main textbooks in the handbook, and made available in the university's library. To reflect contemporary language usage, the selection of the books was kept at the period from year 2000 onwards. These two books are also suggested as textbooks for a few other compulsory subjects. This fact further manifests the importance of the books to the students of the faculty.

The corpus is named as the *Reference Books Corpus* (RBC), with 34 texts, which is actually the total number of chapters from both textbooks. The final size of RBC is 425,854 running words.

British National Corpus (BNC) – The Reference Corpus

This corpus consists of 100 million tokens, which are collected from written and spoken British English. It represents the English used from the 20th century onwards. The written collection makes up 90% of the corpus, and the samples were taken from extracts of newspapers, specialist periodicals and journals, academic books and fictions, published and unpublished letters and memoranda, as well as school and university essays. Ten percent of the corpus, which comprises of the spoken samples, was taken from unscripted informal conversations of volunteers representing various ages, regions and social classes. Apart from that, the samples were also collected from other different contexts, including formal situations, like business and government meetings, to informal situations, like radio shows.

In this study, the BNC acts as a reference corpus to obtain any statistical information on the spread of the lexical patterns exist in the specialised corpora being studied, thus, proving whether the identified patterns are specific to the Engineering English (Meyer, 2002). In other words, BNC serves as the General English, which is used for the comparative study with the E^2C.

Data Analysis Software - Wordsmith 4

The *Wordsmith 4* software is a multi-function software package, which offers programmes for investigating the lexical behaviour in either a single text or a large corpus. It is considered as the best linguistic data analysis software currently available in the market (Someya, 1999), and the "swiss-army knife of lexical analysis" (Sardinha, 1996). This software features the wordlists, keyword, and concordance programmes, which offer various interesting and remarkable tools that are useful for language investigation, such as:

a. ***wordlists*** - the main function of this programme is to generate and maintain alphabetically ordered, frequency ordered or alternative kinds of wordlists, depending on the objectives of a study. A useful information offered by this programme is the statistical details of a studied corpus which include the running words (tokens), types (distinct words), STTR (standardised type token ratio), mean word length, *n*-letter

words etc. This statistical information provides the basic lexical features of the corpus for investigation. The wordlist programme was used in this study to generate the frequency wordlists of the specialised corpus to determine the lexical profiles of the specialised language according to frequency order.

b. ***keyword*** - This programme allows not only a comparison between two wordlist files, but also multiple comparisons; this means, many target files can be analysed against a reference corpus through 'batch processing'. If a word is found to be unusually frequent in a corpus than its frequency in the reference corpus, it is a 'keyword'. Apart from the keyword analysis, another tool used in this study is the key-keyword analysis. The key-keywords are the most frequent keywords in a corpus or any set of files. Therefore, key-key-words are basically the most typical keywords in a corpus (or genre).

c. ***concordance*** - The function of concordance is quite straightforward. It displays the selected words in the contexts it appears as in the original texts. The concordancer is integrated with ***Wordlist*** and ***Keyword*** programmes. The concordance of the target words can be called up directly from these two programmes. This function is most useful to study the behaviour of a lexical unit of interest for its use, meaning and structure by displaying conveniently and clearly the repeated patterns for observation. This programme was used throughout this study, especially when there was a need to observe the neighbouring structures of a word.

METHOD

All the wordlists (frequency, keyword, and key-keyword) of the corpus were generated with the *WordSmith Tools 4.0* software, and the reference corpus employed for the study is the [1]British National Corpus (BNC), which is retrievable from http://www.lexically.net/ downloads/ version4/ downloading%20BNC.htm. Because this study involved a specialised corpus, as much as possible, the texts were kept 'clean', as how they appeared in the original texts. This clean-text policy was proposed by Sinclair (1991) for two reasons. Sinclair propounds that different researchers may set different research aims in corpus data; therefore, the analytical apparatus may cause lack of standardisation, thus, problems for later research of different natures. Another possible issue which can cause difficulties for future research is the lack of standardisation on basic linguistic features such as the identification of words and assignment of morphological division. Therefore, by keeping the texts clean, the potential issues arising from these discrepancies can be minimised.

RESULTS AND DISCUSSION

Table 2 displays the basic statistical details of RBC. The STTR value indicates that there are 27 word types in every 1000 words in RBC. STTR suggests the lexical variation or diversity of the corpus (Banerjee &

Papageorgiou, 2009). A low value means many of the same words are used repeatedly, and a high value suggests the corpus comprises a variety of words, which are less repeated. Therefore, the STTR value suggests that RBC contains many repeated words. This statistical information is especially meaningful if the study involves a comparative analysis of corpora, for example, of different text types or genres.

The following section involves the reporting of the findings (words) according to the types of wordlist and the interpretation of the information (i.e. the lexical features of the specialised language) derived from each wordlist analysis.

TABLE 2
Basic statistical data of RBC

Statistical Details	RBC
tokens used for word list	374,726
types (distinct words)	5,935
standardised TTR	27

RBC Frequency Wordlist Analysis

Table 3 shows the top 50 frequent words in RBC.

The top 50 frequency list shows that the words of higher rank in RBC include a number of function words. As a matter of fact, the top 7 most frequent words in this corpus are function words. The distribution of functions words to content words in the corpus is as illustrated in Fig.1. It shows that function words made up 3% of the words in the corpus, while content words made up 97 %. Out of 5,935 word types in RBC, there are 168 function word types, with 170, 615 tokens.

It should be noted that there are words used repeatedly throughout a text, and these words, therefore, have high frequencies in a language. This also means that the words have high text coverage in that language or corpus. However, the text coverage of the function words in the corpus, as illustrated in Fig.2, suggests that though with smaller number of function words, RBC has a relatively high proportion of function words throughout the corpus, with 46%. In other words, the function words have been used highly repeatedly in RBC. Nevertheless, it should also be noted that the words in this study were classified according to their prototype categories; therefore, there are possibilities that some of the function words identified may not operate as function words in their respective contexts. As such, this finding suggests the frequency and text coverage of the individual word (form), without reference to their functions.

Interestingly, Table 3 also shows that RBC has quite a number of content words in its top 50 list, with approximately half of the list. Those include *voltage, current, circuit* and *output*. This implies a feature of the specialised language, i.e., words of technical nature are frequently used in the corpus; therefore, RBC is not a general English corpus. It can also be seen, with 50 word types, RBC has text coverage of almost half of the corpus (49%). This, once again, highlights the characteristic of RBC, whereby many words are used highly repeatedly in the corpus.

The content words were, subsequently, categorised according to GSL, AWL and

TABLE 3
RBC frequency list (top 50)

	RBC				RBC			
				Cum.				Cum.
N	Word	Freq.	%	%	N	Word	Freq.	%	%
1	THE	40,575	10.83	10.83	26	AT	2,078	0.55	42.79
2	OF	12,966	3.46	14.29	27	FIG	1,996	0.53	43.32
3	IS	12,128	3.24	17.52	28	SIGNAL	1,887	0.5	43.83
4	IN	10,323	2.75	20.28	29	FIGURE	1,874	0.5	44.33
5	A	9,881	2.64	22.92	30	GAIN	1,858	0.5	44.82
6	AND	9,396	2.51	25.42	31	ON	1,597	0.43	45.25
7	TO	7,813	2.08	27.51	32	FROM	1,594	0.43	45.67
8	VOLTAGE	5,181	1.38	28.89	33	OR	1,536	0.41	46.08
9	THAT	4,395	1.17	30.06	34	WHICH	1,447	0.39	46.47
10	FOR	4,344	1.16	31.22	35	SHOWN	1,432	0.38	46.85
11	AS	3,741	1	32.22	36	RESISTANCE	1,417	0.38	47.23
12	BE	3,673	0.98	33.2	37	LOAD	1,268	0.34	47.57
13	CURRENT	3,526	0.94	34.14	38	SOURCE	1,234	0.33	47.9
14	CIRCUIT	3,378	0.9	35.04	39	IF	1,233	0.33	48.23
15	OUTPUT	3,200	0.85	35.9	40	AMPLIFIER	1,201	0.32	48.55
16	ARE	2,775	0.74	36.64	41	THEN	1,181	0.32	48.86
17	WE	2,710	0.72	37.36	42	FREQUENCY	1,121	0.3	49.16
18	INPUT	2,515	0.67	38.03	43	DIODE	1,109	0.3	49.46
19	WITH	2,411	0.64	38.68	44	EXAMPLE	1,020	0.27	49.73
20	THIS	2,337	0.62	39.3	45	IT	970	0.26	49.99
21	AN	2,277	0.61	39.91	46	EMITTER	969	0.26	50.25
22	BY	2,274	0.61	40.51	47	CIRCUITS	964	0.26	50.5
23	WILL	2,204	0.59	41.1	48	DC	949	0.25	50.76
24	CAN	2,130	0.57	41.67	49	SMALL	926	0.25	51
25	TRANSISTOR	2,110	0.56	42.23	50	WHEN	917	0.24	51.25

Others. GSL contains a list of 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 and said to be the most frequent English words. AWL, on the other hand, is an academic word list developed by Coxhead (2000). This list stemmed from the needs 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 are found infrequent in non-academic texts. In other words, these words do not appear in GSL. Next, *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, lace 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.

Fig.3 shows the distributions of GSL, AWL and *Others* (word types) in RBC. It appears that RBC has a balanced proportion of *Others* and GSL, that is, 41%. Nevertheless, the text coverage of the word categories provides interesting information. Fig.4 plots the text coverage of the categories for RBC. Within the corpora, it appears that GSL has the highest text coverage (i.e. 77%), followed by AWL (12%) and *Others* (11%). Therefore, the word categories suggest the specialised or technical nature of RBC; however, the text coverage of the word categories implies that general or most frequent English words are still used by the authors to explain the technical concepts in the texts.

The frequency lists thus far reveal the general lexical profile of RBC. It "... offers an ideal starting point for the understanding of a text in terms of its lexis" (Scott & Tribble, 2006). However, looking at the lists, there is a need to determine the significance of the word occurrences in the corpus by comparing them with a reference corpus. Therefore, the following section discusses another type of wordlist analysis - the keyword analysis.

RBC Keyword List Analysis

The total of keywords in RBC is 1,647. The keywords make up 80.9% of the text coverage in the corpus (Table 4). Positive keywords occurred **more** often than would be expected by chance in RBC in comparison with BNC; conversely, negative keywords occurred **less** frequently in RBC than would be expected by chance in comparison with BNC. Table 5 lists both the positive and negative keywords. The negative keywords are reordered from the most negative keywords.

TABLE 4
No. of types and text coverage (%) of RBC keywords

	Types	Text Coverage (%)
Positive Keywords	1193	68.8
Negative Keywords	454	12.1
TOTAL	1647	80.9

Table 5 shows the prevalent use of nouns as most keywords in comparison with the frequency wordlist earlier. In fact, within these 50 most keyed words, there are only two function words included, and these are *is* and *the*. The keyword list highlights the words which are found significant in the corpus; thus, it features the specialised quality the corpus as a collection of technical texts.

Fig.5 provides the distribution of function words to content words in the keyword lists of the corpus. There seems to be an adjustment taking place in the distribution of function words. The

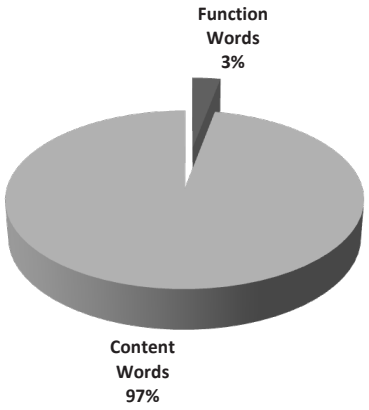

Fig.1: The distribution of function words and content word types in RBC (%)

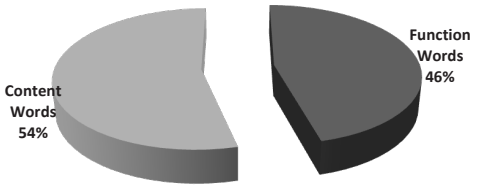

Fig.2: Text coverage of function and content words in RBC

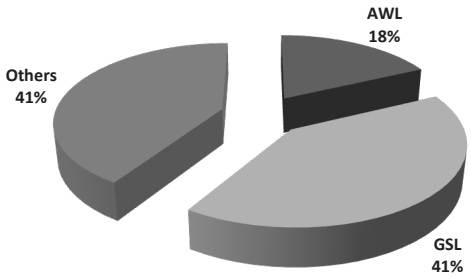

Fig.3: The distribution of GSL, AWL and other word types in RBC

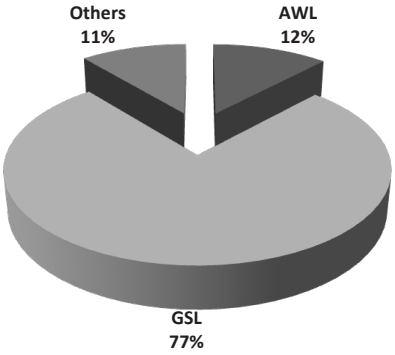

Fig.4: The text coverage of GSL, AWL and Others in RBC

TABLE 5
RBC positive and negative keyword lists (top 50)

N	RBC					
	Positive Keywords			Negative Keywords		
	Keyword	%	Keyness	Keyword	%	Keyness
1	VOLTAGE	1.383	51221.71	I	0.026	-5552.56
2	CIRCUIT	0.901	28558.23	WAS	0.094	-5204.78
3	OUTPUT	0.854	22907.77	HE	0.002	-5118.77
4	TRANSISTOR	0.563	21335.94	YOU	0.026	-4321.62
5	CURRENT	0.941	20788.36	IT	0.259	-3385.17
6	INPUT	0.671	19013.58	HAD	0.006	-3362.90
7	SIGNAL	0.504	13906.71	THEY	0.041	-2266.60
8	GAIN	0.496	12140.65	BUT	0.113	-1765.68
9	FIG	0.533	11886.18	SAID	0.002	-1620.31
10	AMPLIFIER	0.321	11456.36	WHAT	0.011	-1617.95

TABLE 5 (continue)

11	DIODE	0.296	11339.41	WERE	0.051	-1566.64
12	EMITTER	0.259	10212.24	THEIR	0.030	-1485.09
13	IS	3.236	9696.60	TO	2.085	-1285.46
14	RESISTANCE	0.378	9439.86	THEM	0.006	-1265.69
15	LOAD	0.338	8554.33	BEEN	0.048	-1250.84
16	CIRCUITS	0.257	8365.97	ME	0.001	-1094.86
17	FIGURE	0.500	8272.17	OUT	0.028	-1078.71
18	FREQUENCY	0.299	7518.02	YOUR	0.003	-1065.90
19	TRANSISTORS	0.199	7428.06	DO	0.022	-1012.04
20	THE	10.828	6854.70	UP	0.038	-1007.74
21	DC	0.253	6658.85	ON	0.426	-991.37
22	BIAS	0.224	6183.92	HAVE	0.210	-914.08
23	COLLECTOR	0.211	6163.08	LIKE	0.018	-851.72
24	SHOWN	0.382	5998.61	NOT	0.207	-845.74
25	SOURCE	0.329	5952.75	THERE	0.103	-825.05
26	OP	0.183	5827.19	KNOW	0.010	-769.70
27	FEEDBACK	0.200	5468.72	DON'T	0.001	-765.08
28	BIASED	0.174	5435.93	THINK	0.001	-740.42
29	CONFIGURATION	0.185	5222.92	DID	0.002	-718.67
30	AC	0.175	5092.65	ALL	0.111	-708.38
31	CAPACITOR	0.140	4819.57	ABOUT	0.053	-706.25
32	RESISTOR	0.134	4815.36	COULD	0.026	-693.10
33	SATURATION	0.140	4695.39	NO	0.079	-691.76
34	DEVICE	0.205	4621.08	WELL	0.029	-652.54
35	MOSFET	0.115	4503.73	WOULD	0.084	-649.46
36	VOLTAGES	0.121	4448.34	GET	0.008	-627.02
37	CHARACTERISTICS	0.207	4307.69	YEARS	0.006	-594.63
38	PARAMETERS	0.151	4019.18	WORK	0.007	-588.70
39	LOOP	0.147	3992.94	NEW	0.024	-588.01
40	EQUIVALENT	0.200	3904.25	AFTER	0.019	-571.12
41	IMPEDANCE	0.117	3888.38	WAY	0.012	-552.10
42	BASE	0.240	3806.71	JUST	0.028	-544.01
43	GATE	0.183	3787.06	OWN	0.002	-529.26
44	CURRENTS	0.125	3519.06	GO	0.012	-476.48
45	REGION	0.220	3326.66	OVER	0.037	-471.79
46	DRAIN	0.130	3317.77	COME	0.004	-470.13
47	BIPOLAR	0.083	3005.36	SAY	0.006	-435.02
48	NETWORK	0.189	2992.84	DAY	0.003	-430.96
49	CAPACITANCE	0.086	2937.97	WORLD	0.001	-429.81
50	ZERO	0.137	2930.95	LIFE	0.001	-427.79

occurrence of function words is more in the keyword list than its occurrence in the frequency wordlist (see Fig.1). RBC has about 7% of function words in the keyword list in comparison with 3% in the frequency wordlist. Table 6 provides details of the positive and negative key-function-words in terms of the number of types and text coverage.

TABLE 6
No. of types and text coverage (%) of RBC key-function-words

	Types	Text Coverage (%)
Positive Key-function-words	29	22.2
Negative Key-function-words	85	12.9
TOTAL	114	35.1

Table 7 shows that the negative key-function-words of RBC are mostly pronouns such as *I*, *he*, *you*, *they*, *them* and *me*. Though *was* and *had* need further distinctions in terms of their auxiliary-verb functions, both indicate that tenses may also be a characteristic that distinguishes RBC from general English texts. The addition of words like *were* and *did* further implies that the use of past tenses is not likely characterising the corpus. Function words appear to be at least the most 20 negative keyed words in the corpus. Delexicalised verbs (*get*, *know*, *got*) in the corpus further mark the sharp contrast between RBC and general English texts. The same idea is projected by other general words in the negative keyword lists such as *people*, *life*, *world*, and *day*.

The distribution of GSL, AWL and *Others* categories is as presented in Figure 6. The comparison between the distributions of all the word categories in the frequency wordlist and keyword list shows a noticeable difference. The keyness notion highlights the use of GSL and slightly reduces *Others*. Table 8 gives the comparison of the distributions. Though the difference in the distributions is relatively small between the two lists, the keyword list proves the significance of the word occurrence.

TABLE 7
Positive and negative key-function-words of RBC

RBC					
Positive Key-function-words			Negative Key-function-words		
Key word	%	Keyness	Key word	%	Keyness
IS	3.24	9696.60	I	0.74	-5552.56
THE	10.83	6854.70	WAS	0.87	-5204.78
CAN	0.57	1141.01	HE	0.60	-5118.77
WE	0.72	1121.98	YOU	0.59	-4321.62
SINCE	0.23	1045.35	IT	0.93	-3385.17
WILL	0.59	853.27	HAD	0.42	-3362.90
VERSUS	0.04	667.68	THEY	0.38	-2266.60
ACROSS	0.12	564.54	BUT	0.45	-1765.68

TABLE 7 (continue)

AN	0.61	363.73	WHAT	0.23	-1617.95
IN	2.75	361.93	WERE	0.31	-1566.64
WHEREAS	0.05	336.17	THEIR	0.26	-1485.09
THEN	0.32	303.51	TO	2.61	-1285.46
ARE	0.74	261.98	THEM	0.17	-1265.69
AS	1.00	241.79	BEEN	0.26	-1250.84
BETWEEN	0.20	239.79	ME	0.13	-1094.86
HOWEVER	0.14	223.29	OUT	0.20	-1078.71
BE	0.98	216.15	YOUR	0.14	-1065.90
EACH	0.14	209.77	DO	0.18	-1012.04
OPPOSITE	0.03	160.48	UP	0.21	-1007.74
MUST	0.13	110.92	ON	0.74	-991.37
OFF	0.13	104.04	HAVE	0.45	-914.08
THAN	0.24	96.62	LIKE	0.15	-851.72
ABOVE	0.06	90.80	THERE	0.29	-825.05
TOWARD	0.01	89.82	DON'T	0.09	-765.08
THROUGH	0.14	76.96	DID	0.09	-718.67
FOR	1.16	59.15	ABOUT	0.19	-706.25
MINUS	0.01	58.84	COULD	0.14	-693.10
THIS	0.62	57.23	NO	0.23	-691.76
BELOW	0.03	49.35	WOULD	0.23	-649.46

TABLE 8
A comparisons of the distribution between frequency and keyword lists

	Frequency	
	Word List	Key Word List
GSL	41	48
AWL	18	15
Others	41	37

RBC Key-Keyword List Analysis

Table 9 gives the first 100 key-keywords from RBC. The key-keyword list of RBC consists of 589 word types. This top 100 list shows a reasonable coverage of the main word classes, apart from the predominant nouns such as verbs and adjectives. Verbs, such as *shown*, *determine*, *connected*, *applied*, *shows*, *using, assume*, *analyse*, *defined*, *determined*, *calculate*, *obtain*, and *consider*, and adjectives, such as *equivalent*, *negative, constant*, *basic*, *positive*, and *equal*, are listed in this top 100 of key-keyword list. It appears that in RBC, the use of adjectives is to highlight the standard concepts in its description such as the words *equivalent*, *constant*, *basic*, and *equal*. The abbreviations and symbols such as *DC*, *AC*, *B*, *PSPICE*, *fig*, *EQ* and *V* are also included in the list.

It appears that more function words (*is*, *the*, *can* and *we*) are included in this list. The whole proportion of the function words is as shown in Fig.7. It shows that there are 24 word types (4%) identified as the function

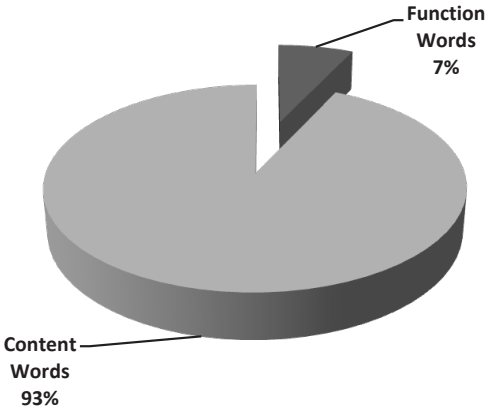

Fig.5: The distribution of function and content words in the RBC keyword list (%)

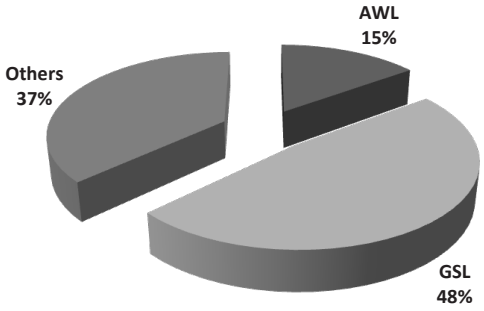

Fig.6: The distribution of GSL, AWL and Others word types in RBC (%)

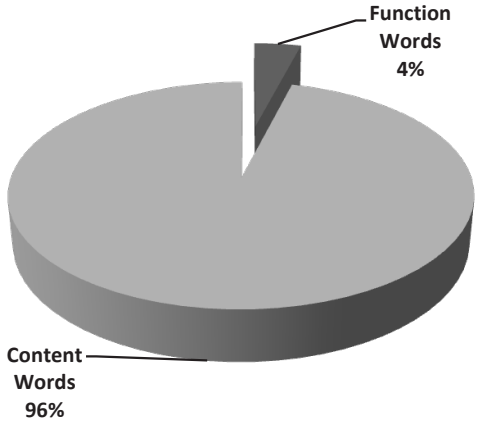

Fig.7: The distribution of function and content words in the RBC key-keyword list (%)

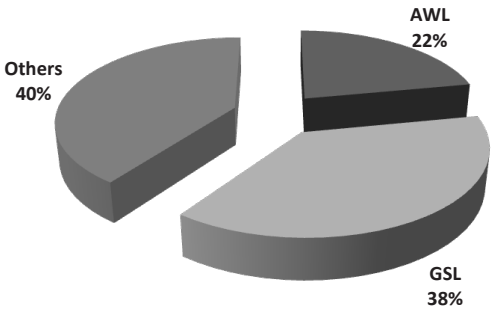

Fig.8: The distribution of GSL, AWL and other word types in RBC (%)

words in RBC. The list of all the key-key-function words in the corpus is presented in Table 10. There are two modals available in the list: *can*, and *will*. An interesting discovery is the pronoun *we*; the pronoun is listed as one of the highest ranked words in the frequency list, and it is still one of the key-keywords. The key-keyword analysis proves the significance of the pronoun in the specialised corpus. Another pronoun in the list is *each.* Meanwhile, *for*, *between*, *above*, *across*, *in*, *off, versus* and *through* are the prepositions included in the list. Also in the lists are conjunctions like *since*, *as*, *whereas*, *then* and *however*, and determiners, *a* and *the*. Briefly, these key-keyword lists provide a set of words of wider range in terms of word class category.

The distribution of GSL, AWL and *Others* categories can be observed from Fig.8. The *Others* category has the most number of word types, followed by GSL and

TABLE 9
Key-keyword lists of RBC (top 100)

N	KW	Texts	%	Overall Freq.	N	KW	Texts	%	Overall Freq.
1	CIRCUIT	34	100	3378	51	CONFIGURATION	21	61	680
2	IS	34	100	12128	52	CURRENTS	21	61	450
3	SHOWN	34	100	1432	53	AMPLIFIERS	20	58	192
4	VOLTAGE	34	100	5181	54	B	20	58	733
5	CURRENT	33	97	3519	55	FREQUENCY	20	58	1087
6	THE	33	97	40358	56	PSPICE	20	58	132
7	INPUT	32	94	2508	57	CHAPTER	19	55	402
8	TRANSISTOR	32	94	2110	58	EQUATIONS	19	55	237
9	DC	31	91	936	59	MAXIMUM	19	55	452
10	OUTPUT	31	91	3186	60	SATURATION	19	55	507
11	RESISTOR	31	91	496	61	TERMINALS	19	55	192
12	SIGNAL	31	91	1878	62	CONSTANT	18	52	324
13	VALUE	31	91	852	63	REGION	18	52	767
14	CIRCUITS	30	88	954	64	ASSUME	17	50	306
15	RESISTANCE	30	88	1408	65	BASE	17	50	860
16	VOLTAGES	30	88	451	66	BASIC	17	50	364
17	BIAS	28	82	825	67	BIPOLAR	17	50	301
18	CHARACTERISTICS	27	79	762	68	FIG	17	50	1996
19	EXAMPLE	27	79	974	69	JUNCTION	17	50	414
20	MAGNITUDE	26	76	369	70	OBJECTIVE	17	50	241
21	SOLUTION	26	76	487	71	ANALYZE	16	47	84
22	AC	25	73	648	72	COLLECTOR	16	47	770
23	AMPLIFIER	25	73	1192	73	COMMENT	16	47	223
24	DETERMINE	25	73	532	74	DEFINED	16	47	365
25	EMITTER	25	73	967	75	DESIGN	16	47	499

TABLE 9 (continue)

26	LOAD	25	73	1249	76	DETERMINED	16	47	383
27	PARAMETERS	25	73	551	77	PARAMETER	16	47	212
28	CAPACITOR	24	70	513	78	QUIESCENT	16	47	164
29	CONNECTED	24	70	338	79	RATIO	16	47	176
30	FIGURE	24	70	1795	80	SIMULATION	16	47	120
31	RESISTORS	24	70	178	81	SINUSOIDAL	16	47	147
32	SOURCE	24	70	1191	82	WE	16	47	2099
33	VALUES	24	70	434	83	BIASING	15	44	128
34	ANALYSIS	23	67	768	84	CALCULATE	15	44	149
35	APPLIED	23	67	631	85	CAPACITANCE	15	44	305
36	BIASED	23	67	645	86	CONFIGURATIONS	15	44	163
37	CAN	23	67	1750	87	DEVICES	15	44	238
38	DEVICE	23	67	740	88	EQ	15	44	276
39	EQUIVALENT	23	67	717	89	MOSFET	15	44	429
40	RESULTING	23	67	437	90	OBTAIN	15	44	167
41	SHOWS	23	67	616	91	POLARITY	15	44	126
42	TRANSISTORS	23	67	735	92	POSITIVE	15	44	311
43	USING	23	67	694	93	REVERSE	15	44	363
44	ZERO	23	67	480	94	SIGNALS	15	44	213
45	DIODE	22	64	1103	95	THEREFORE	15	44	367
46	GAIN	22	64	1839	96	V	15	44	855
47	IMPEDANCE	22	64	421	97	CONSIDER	14	41	267
48	NEGATIVE	22	64	386	98	EQUAL	14	41	194
49	SINCE	22	64	707	99	FOLLOWER	14	41	225
50	TERMINAL	22	64	383	100	FUNCTION	14	41	461

AWL, while in the previous two wordlists, the order is GSL, *Others* and AWL. This means that the key-keyword analysis has further enhanced the specialised quality of the corpus. The information obtained from the key-keyword analysis is very useful for the selection of words for the researcher's further study because the distributions of the categories in this key-keyword list is regarded as sufficient to supply words for further analysis, taking into consideration the range covered by the words.

The frequency wordlist is able to highlights the specific features of RBC as a specialised language, in this case, technical texts. Although the frequency list shows the use of a number of function words as among the most frequent words in the corpus, there also seems to be a balanced use of high frequency content words. The coverage of the first 50 high frequency word in the corpus, i.e. 49%, suggests lesser use of words, thus, underlines the specialised quality of the RBC texts.

Next, the keyword list provides different but more detailed features of the corpus. It exposes more specific and technical words in the corpus. The analysis of the positive and negative keywords further distinguishes the specialised language in comparison with the general English texts (BNC). There are more technical nouns appear in the higher rank of the list, suggesting the technical concepts available in the specialised language. The negative keywords reveal more about the language. Nelson (2000) notes that it is possible to describe the language in a specific domain by investigating 'what is *not* found there'. This can be achieved by using the negative keywords. Apparently, pronouns, past tenses and delexicalised verbs occur less frequently in RBC, in comparison to BNC.

The key-keyword list offers more varied members in the top list. The list comprises of lesser numbe of words but still sees 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 RBC – the study embarked by the researcher at a later stage.

With reference to the distribution of function words, the analyses of the three wordlists show that the frequency wordlist has the lowest distribution of function words, while the keyword list has the highest distribution. Despite the fact that the distribution in the keyword lists 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 function words which are significant and occur in more than 2 texts in each corpus (this parameter was set at the beginning of the analysis). Furthermore, unlike frequency and key word lists, the higher ranked key-keywords cover a wide range of function words including prepositions, pronouns, conjunctions, modals and other auxiliary verbs.

With regard to the distributions of GSL, AWL and Others categories, the

analyses of the three wordlists prove that the specialised corpus has different proportions of vocabulary types from general English. As proposed by Nation (2001), high frequency words (GSL) constitute 80% of the tokens in a text (corpus), while academic words (AWL) make up 9%, and 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). Meanwhile, the prevalence of the *Others* category in the RBC key-keyword list is accounted by the consistency of word occurrence in a number of texts across the corpus, which is the main quality of the key-keyword list. Though the total proportions are different from one word list to another, the pattern is rather similar. The key-keyword analysis not only highlights significant words in the corpus, but also identifies the consistency of the word occurrence across the corpus (range).

The investigation in this paper provides a preliminary finding for the description of the specialised language, RBC, particularly from the perspectives of the distributions of function and content words, as well as the GSL, AWL and *Others* categories. The proven fact is that while frequency wordlists provide information on the lexical foundation of a text or corpus, keyword lists result in the identification of significant words, which inform what a text or a corpus is about (aboutness). Key-keyword lists, on the other hand, inform the range of the keywords in terms of the number of texts they appear in a corpus; the more texts a key word occurs in, the more 'key-key' it is. All these lists prove to offer a cornucopia of information related to language use, depending on the predetermined objectives in a language study. This paper does not only delve into the differences present between these three types of list, but also the similarities.

TABLE 10
Key-key-function-words of RBC

A	OFF
ABOVE	SINCE
ACROSS	THE
AN	THEN
ARE	THIS
AS	THROUGH
BE	VERSUS
BETWEEN	WE
CAN	WHEREAS
EACH	WILL
FOR	HOWEVER
IN	IS

CONCLUSION

This paper underlines the useful application of the different types of wordlists analysis, namely, the Frequency Wordlist, Keyword List, and Key-keyword List, in highlighting the lexical profiles of a specialised language. The results from the word list analyses reveal that different aspects of the language show up in all the wordlists. The findings also revealed that despite the fact that the distributions of function words, GSL, AWL and *Others* categories vary from one list to another, these categories are generally retained in all the lists with small differences.

ACKNOWLEDGEMENTS

This work stemmed from a PhD study, which was sponsored by the Ministry of Higher Education, with the support from UTeM. The PhD study is currently pursued in UKM, Bangi.

REFERENCES

Baker, M. (1988). Sub technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language, 4*(2), 91-105.

Banerjee, J., & Papageorgiou, S. (2009). *Analysing written language*. [PowerPoint slides].

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103-116.

Coxhead, A. (2000). The Academic Word List: A corpus-based word list for academic purposes. In B. Kettemann & G. Marko (Eds.), *The Fourth International Conference on Teaching and Language Corpora* (pp. 73-89). Armsterdam: Rodopi.

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education, 10*, 127-143.

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

Granger, S., & Paquot, M. (2009). *In search of a General Academic vocabulary: A corpus-driven study.* Paper presented at the International Conference on L.S.P, Heraklion, Crete.

Hunston, S. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Philadelphia: John Benjamins Publishing Co.

Martı´nez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes 28(3), 183-198.*

Martin, A. V. (1976). Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly, 10*(1), 91-98.

Menon, S., & Mukundan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. *Pertanika Journal of Social Science and Humanities 18*(2), 241 – 258.

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, 25,* 235-256.

Mukundan, J., & Aziz, A. (2009). Loading and distribution of the 2000 high frequency words in Malaysian English language textbooks for form 1 to form 5. *Pertanika Journal of Social Science and Humanities 17*(2), 141 – 152.

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

Nelson, M. (2000). *A Corpus-based Study of the Lexis of Business English and Business English Teaching Materials* (Unpublished doctoral dissertation). University of Manchester, UK.

Noorzan Haji Mohd Noor. (2005). *Corpus-based research for professional communication: The real perspective for language education*. Paper presented at the Borneo Language Teaching Conference (BLTC), Promenade Hotel, Kota Kinabalu.

Perez-Paredes, P. (2003). Small corpora as assisting tools in the teaching of English news language: A preliminary tokens-based examination of Michaels Swan's practical English usage news language wordlist. *English for Specific Purposes World, 2*(3). Retrieved from http://www.esp-world.info/articles_6/pascual.htm

Sardinha, B. (1996). Review: Wordsmith tools. *Computers & Texts, 12*. Retrieved from http://users.ox.ac.uk/~ctitext2/publish/comtxt/ct12/sardinha.html

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins Publishing Company.

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

Someya, Y. (1999). *A corpus-based study of lexical and grammatical features of written business English.* (Master's thesis). Graduate Department of Language and Information Sciences, University of Tokyo.

Tsubaki, M. (2004). Vocabulary in English for academic purposes: A corpus study of journal article. *Bunkyo Gakuin Junior College Bulletin 4,* 159-168. Retrieved from http://cicero.u-bunkyo.ac.jp/lib/kiyo/fsell2004/159-168.pdf

Worthington, D., & Nation, I. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal 27*(2), 1-11.

ENDNOTE

[1] The reference corpus employed to generate the keyword lists in this study is the British National Corpus (BNC). This corpus consists of 100 million tokens, which are collected from written and spoken British English. It represents the English used from the 20th century onwards. The licence of this modern mega-corpus can be easily obtained online at http://bncweb.info/. BNC serves as the general English, which is used for the comparative study with the specialised language, RBC.