3L: The Southeast Asian Journal of English Language Studies – Vol 21(1): 117 – 131
http://dx.doi.org/10.17576/3L-2015-2101-11

Exploring Word Associations in Academic Engineering Texts

NOORLI KHAMIS
Centre for Languages and Human Development (PBPI)
Universiti Teknikal Malaysia Melaka
Malaysia
noorli@utem.edu.my

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

ABSTRACT

Given the importance of lexis in language description, this study attempts to integrate the lexical approach to describe a specialised language for teaching and learning. In addition, this paper demonstrates the use of the correspondence analysis (CA), one of the multivariate techniques, as a useful tool to describe a language. As such, this is a corpus-based study of verbs among academic engineering text types. A larger engineering corpus (E^2C) was constructed by combining two specialised corpora, consisting of two text types, namely reference books (RBC) and journal articles (EJC). The Wordsmith 6 program was used to extract 30 key-key-verbs from E^2C. The British National Corpus (BNC) was used as the reference corpus. The CA was conducted with these key-key-verbs by computing the frequency values of the verbs generated for each corpus: E^2C, RBC, EJC and BNC. The findings include the visual display of the complex inter-relationship of the verbs among the corpora, thus, demonstrating the potential use of the CA as a tool for specialised language description. The empirical observations of the verbs may lead to significant findings on the features of the academic engineering texts types; thus, this study promises more well-informed future investigations into other linguistic features, rhetorical functions, and pedagogical implications involving the academic engineering texts.

Keywords: correspondence analysis; academic engineering texts; verbs; corpus-based study; specialised corpora

INTRODUCTION

In many English for Specific Purposes (ESP) and English for Academic Purposes (EAP) classrooms, the construction of specialised corpora has been carried out with the aim to sample the language that is considered to be the central concern of learners' needs. Examples of such corpora include the Guangzhou Petroleum English Corpus (GPEC), the Hong Kong University of Science and Technology (HKUST) Computer Science Corpus, the Jiaotong Diaxue English of Science and Technology (JDEST) Corpus and the Student Engineering English Corpus (SEEC). Gavioli (2005) points out that unlike other general English corpora, such as the British National Corpus (BNC), Brown Corpus, American National Corpus (ANC) and ICAME corpora collection, the construction of such specialised corpora offers some distinct advantages. One of them is being highly controlled in the quantity of words collected, which results in manageability from the technical aspect, as well as the analytical point of view. Another advantage of the specialised corpus design involves the features of the specialised language, of which offer information about the language ad hoc to the learners' needs.

In linguistic studies, the application of electronic corpora promises access to a specific domain linguistic knowledge, including the instances of how a particular lexical item is used in its context. Hence, corpus-based data provide great assistance to many researchers to exhaustively explore the linguistic features of ESP/EAP from all the possible sources of written and spoken texts – books, magazines, websites, CDs, interviews, plays, etc. Analyses of corpus-based data allow researchers to highlight recurrent features of a language, making it possible to provide a more substantiated procedure to describe the humdrum and routine aspects of ESP/EAP, which have been the concern among the teachers who have to assume the roles of experts for both language and the specific discipline (Fadhil 2001). It offers a means to isolate and provide "indications about key lexical, grammatical or textual issues to deal with in ESP classes" (Gavioli 2005). Thus, the integration of corpora in ESP/EAP is "viewed as a coherent course design step at university settings" (Fuentes & Rokowski 2003).

THE LEXICAL APPROACH TO LANGUAGE DESCRIPTION

According to Lewis (1993), language is made up of 'grammaticalised lexis' instead of 'lexicalised grammar'. This simply means that meanings are constructed from fixed words rather than fixed structures. Therefore, the basis of a language is its lexis. Mudraya (2006) draws a distinction between vocabulary and lexis as the former being a collection of individual words with meanings and the latter being a combination of words (not just single words) stored in our mental lexicons to be used anytime. In addition, Harwood (2002) sees lexis as strings of words which go together, and this includes prefabs and collocations. Harwood also affirms that it is quite difficult to discriminate between lexis and the traditional concept of 'grammar'. This vagueness suggests the importance and priority of lexis in language description.

Lexical items of a language are regarded as socially sanctioned independent units, of which many are words, and in fact more are multi-word units. Lewis (1993) suggests the following taxonomy of lexical items:

a) words – dictionary entries
b) polywords – two to three words which operate like an individual lexical items, for example, phrasal verbs.
c) collocations – individual words with its neighbouring words
d) institutionalised expressions – fixed items (*not yet*, *certainly not*, *sorry to interrupt, but can I just say…..*)
e) full sentences – readily identifiable pragmatic meaning.

Lewis also maintains that language mastery involves assembling lexical units from the smallest components. Nattinger and DeCarrico (1992, p. xv) note that:

> One common pattern in language acquisition is that learners pass through a stage in which they use a large number of unanalysed chunks of language in certain predictable social contexts. They use, in other words, a great deal of 'prefabricated' language. Many early researchers thought these prefabricated chunks were distinct and somewhat peripheral to the main body of a language, but more recent research puts this formulaic speech at the very centre of language acquisition and sees it as basic to the creative rule-forming processes which follow.

Therefore, the lexical approach proposes that if learners are provided with chunks of language according to its particular context, they will be able to master these chunks, which later become the raw data by which learners perceive patterns of language or syntax.

The existence and importance of these language chunks has been the central concern of many linguists. Sinclair (1991) suggests the idiom principle in discussing collocation, and claims that words do not occur at random in text. Martinez and Schmitt (2012) assert that lexical phrases play an important role for fluent language production and vocabulary learning. Menon and Mukundan (2012) also stress that our knowledge of a language involves not only individual words and phrases, but also co-occuring words in a cohesive text. Firooz et al. (2012) underscores the relationship between forms and functions of formulaic strings for pedagogical purpose. There are more existing and on-going investigations on language chunks, which are also variably referred as lexical phrases (Nattinger & DeCarrico 1992), collocations (Lewis 1993, Menon & Mukundan 2012), multi-words items or units (Monti 2011), and formulaic sequence.

The identification and presentation of these language chunks has become the central task of many researchers into language descriptions, and language teaching and learning. In fact, as emphasised by Lewis "…it applies mutatis mutandis to both spoken and written language, and to both ESP and general language" (1993, p. 96).

Given the importance of lexis in language description, this study attempts to integrate the lexical approach to describe language for teaching and learning. The focus on words is essential because it provides the basis for the investigation of the specialised language, i.e. the academic engineering texts.

SPECIALISED CORPORA AND GENRE ANALYSIS

Since the ascendancy of computational data analysis in language description, there have been continuous ventures into many potential applications of corpus work in ESP/EAP, which include variations across genres, genre conventions and needs analysis (Gavioli 2005). Lee (2001) also notes that many corpus-based investigations into language description involve genres or related concepts such as registers, text-types, domain, style, sublanguage, message form and the like. With the potential offered by specialised corpora, genre studies are becoming more extensive and interesting. Studies such as comparisons between corpora of texts from different genres, such as Business English published materials and Business English Corpus (of 28 macro-genres of spoken and written Business English), are able to highlight the specific characteristics of both corpora (Nelson 2000).

On the same note, the need to study the linguistic features of different genres (or text types) in a language is deemed essential in describing the nature of the language, which apparently has distinct and significant differences from one genre to another (Fadhil 2001). There have been many studies into ESP/EAP courses, aimed to identify and describe the linguistic features of specialised languages, including the lexis (Noorli & Imran Ho 2012), rhetorical structures (Kanoksilapatham 2013), lexico-grammatical patterns (Menon & Mukundan 2012), discourse, and other aspects of linguistic descriptions. Different syntactic patterns or moves in different sections of article can be observed from a comparative study between sections in article writing. By comparing a specialised corpus with a general corpus, a variety of rhetorical patterns can be identified. The potential of specialised corpora to capture local characteristics of a language, such as features of genres, promotes corpus work to complement genre studies. In addition, there have also been many authors relating their ideas and experience in designing and preparing ESP/EAP courses and materials (Smith et al. 2007). Many of these findings have been helpful in guiding language instructors to plan and prepare their teaching effectively.

The starting point of an investigation for a language for teaching or/and learning interest is the keywords and their surrounding company, simply because communication is not possible without words (Menon & Mukundan 2012). Jin et al. (2012) assert that the vocabulary component needs to be given due attention in a language course. A good knowledge of the vocabulary allows language instructors and course designers to make informed decision in selecting words that deserve to be given attention in designing a course. As such, an underlying tenet of this study is that the lexical items of a specialised language, in particular the lexical and grammatical properties of that specialised academic genre, must be the focal point in ESP/EAP. For the purpose of this paper, the initial investigation is done with the lexical properties of the academic engineering texts, particularly in exploring the individual words.

MULTIVARIATE STATISTICS FOR GENRE ANALYSIS

More recent genre-based corpus studies have been refined with the application of various statistical measurements. More detailed linguistic descriptions on the features of a language can be gained. Abney (1996) underscores the relevancy of statistical methods in approaching a language description including solving issues of ambiguities, naturalness, grammatical degrees, structural preferences, and error tolerance. In addition to the mutual information (MI), log-likelihood (LL), and chi-square with Yates's correction (Yates) which are built in the Wordsmith program, other statistical techniques have been employed to facilitate a more accurate language description. Chujo et al. (2010), for example, examined a range of statistical measures to identify technical vocabulary, which include LL, MI, Chi2, Yates, the Dice coefficient (Dice), Cosine, complimentary similarity measures (CSM), and frequency. The examination reveals that different statistical techniques can effectively identify specialised vocabulary ranked according to different proficiency levels. In one part of a study, Nishina (2007) presented a comparative study of texts of different genres in the light of multivariate analysis (correspondence analysis, principle component analysis and cluster analysis). The study shows that the empirical analysis facilitates the identification of the internal criteria (similarities, differences and styles) of text genres based on the external data. Using correspondence analysis, Imran Ho and Laman (1997) demonstrate the relationship of top 30 most frequent words in different varieties of English: Brown Corpus, LOB Corpus, New Zealand Newspapers Corpus and Malaysian Newspapers Corpus. With similar approach, Imran Ho (2009) compared the word frequencies in manifestos of different political groups in the 2008 12th Malaysian General Election. The studies reveal words that highly characterise each corpus. Evidently, there is a huge body of mathematical techniques that can be explored in relation to language study to discover what have been previously regarded as complex.

The application of these statistical approaches have allowed the investigations on any similar or differing aspects of genres and text types in a language. The complex links or dispersions between more than two corpora can be examined. The employment of both corpus-driven methods and statistical measurements, such as the multivariate statistics, makes the exploration into these complex links possible.

Multivariate statistical analysis refers to multiple advanced techniques for investigating relationships among multiple variables at the same time. Other commonly encountered multivariate analyses in linguistics research include factor analysis and multidimensional scaling. The underlying assumption is that different kinds of text (corpus) differ in their functions at the linguistic level. It has been discovered that multivariate analyses are a potential statistical analysis to quantify similarities and differences among text types by picturing the relationships visually for further corpus-driven investigation and interpretation. According to McEnery and Wilson (2001), the common aim of all these multivariate analyses is "to summarise a large set of variables in terms of a smaller set on the basis of statistical similarities between the original variables, whilst at the same time losing the minimal amount of information about their differences".

This paper demonstrates the use of correspondence analysis (CA), one of the multivariate techniques, to generate possible visual associations of verbs between the academic engineering corpora and a reference corpus. The empirical observations of the verbs may lead to significant findings on the features of the academic engineering texts types; thus, this study promises more well-informed future investigations into other linguistic features, rhetorical functions, and pedagogical implications involving the academic engineering texts.

METHODOLOGY

CORPORA

Despite the advantages in the application of a specialised corpora for ESP language description as mentioned in the previous section, Gavioli (2005) argues that there is the other side of the coin; the small number of words in the corpora makes such corpora dubious in distinguishing the characterising features of the specialised language and/or making generalisation about the features inside and even beyond the ESP field. This means that although these corpora can provide samples of technical lexis, it does not warrant that the lexis is in fact the features of the language represented in the corpus. This is because a genre and general language are not different categories. Furthermore, it has also been argued that what seems to be relevant to the students, may not coincide with the actual learning needs in the teaching/learning environments. Hence, it is necessary to compare the frequent features of the specialised corpus with:

a) the frequent features of other genres (or text types), and
b) those of the general language

to determine the typical features of the studied corpus. As such, there are two main corpora in this study. The corpora are the:

a) Academic Engineering Texts Corpus (E^2C)
   i) engineering Reference Books Corpus (RBC)
   ii) engineering Journal Articles Corpus (EJC)
b) British National Corpus (BNC)

The E^2C is generally a combination of RBC and EJC, which acts as a general corpus of the academic engineering texts. This corpus consists of 102 texts with a total of 601,481 running words. The engineering discipline of which the texts were collected is the Electrical and Electronics Engineering.

The RBC consists of two reference books for the discipline. This corpus consists of 34 files, which are actually the total of chapters from both reference books. The final size of RBC is

http://dx.doi.org/10.17576/3L-2015-2101-11

374,726 tokens. The EJC is a collection of the engineering journal articles, randomly retrieved from the online databases of a local university. The journals were selected based on the titles of chapters in the reference books. This corpus contains 226,755 tokens. There are 68 files in this corpus.

The BNC acts as a reference corpus to obtain any statistical information on the spread of the lexical patterns exist in the specialised language being studied. In addition, the statistical data inform the probability of the identified patterns as being specific to the engineering English, instead of the general English (Noorli & Imran Ho 2012). This corpus contains 100 million tokens, which are collected from written and spoken British English. It represents the English used from the 20th century onwards.

CORRESPONDENCE ANALYSIS (CA)

The following procedure involves the application of multivariate analysis to discover any complex links or dispersions of verbs between the corpora. The technique employed in the study is correspondence analysis (CA), which enables visual displays of any possible relationships. The analysis is computed based on values derived by cross-tabulating the frequencies of words (verbs) across the corpora. Hence, the initial stage involves generating the different types of wordlists for each corpus for the selection of verbs.

SELECTING THE VERBS

The Wordsmith 6 program was used to generate the frequency wordlist, keywords list and key-keyword list for each engineering corpus: EJC, RBC and E^{2}C. Because different wordlists result in different set of words, the selected verbs for the CA need to come from only one type of wordlist (either the frequency wordlist, keywords list and key-keyword list) to ensure consistency and relevancy for the interpretation of the findings. Therefore, it was determined that the selected verbs to be analysed for the CA were obtained from the E^{2}C key-keyword list since the corpus comprises RBC and EJC. The key-keywords are the most frequent keywords in a corpus or any set of files. Therefore, the employment of key-keywords for this study indicates the key-keyness of the word in the academic engineering corpus; the more texts a word is 'key' in, the more 'key-key' it is. The selection of words from the key-keyword list further substantiates the significance of the words in the specialised language.

This study resorted to the use of (key)word forms, instead of lemmas, because the researcher intends to identify all the significantly frequent word forms used in the corpus for possible further analyses. 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).

There have been many arguments pertaining to the use of word forms versus lemmas for analysis in language description. The head words in dictionary entries are actually 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 (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. 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).

30 verbs were selected from the E^2C key-keyword list. Because there are issues of key-keywords which may take in more than one word class, for example *signal* may be either a noun or a verb, therefore, the selection of these 30 verbs was based on the prototype category, regardless the forms the verb assumes. This means that the word *signal* is regarded as a noun, instead of a verb. In addition, inflectional forms of verbs such as *-s*, *-ed*, and *–ing* in *shows*, *connected* and *using* are also regarded as verbs for analyses.

ANALYTICAL PROCEDURE

Because the CA requires the frequency values of the verbs for each corpus, the frequencies of the same words were obtained from EJC, RBC and BNC frequency wordlists. Table 1 lists the 30 verbs for all the corpora. Unlike other techniques of multivariate statistics, correspondence analysis can be computed with raw frequency data. The CA is computed based on values derived by cross-tabulating the frequencies of the verbs across the corpora. The purpose is to show the statistical similarities and differences across the corpora. The CA procedure was carried out with the XLSTAT 2013 program.

TABLE 1. Frequency of 30 verbs for the CA

Verbs	E^2C	BNC	EJC	RBC
SHOWN	1,756	14,880	324	1,432
SHOWS	863	11,571	217	646
CONNECTED	488	3,301	108	380
DETERMINE	639	3,900	77	562
USING	1,148	24,434	329	819
APPLIED	775	7,549	99	676
DEFINED	498	5,866	75	423
ASSUME	369	4,054	19	350
OBTAINED	397	6,259	147	250
DETERMINED	564	7,575	86	478
CALCULATE	192	1031	13	179
NOTE	500	10,436	86	414
OBTAIN	291	4,551	56	235
USED	1,256	65,980	554	702
ANALYZE	111	117	15	96

CALCULATED	179	2630	54	125
CONSIDER	376	11,590	31	345
INCREASES	349	4,232	64	285
PROVIDES	342	8,354	125	217
CONTROLLED	221	4584	89	132
DECREASES	168	403	25	143
RESULT	618	21,938	130	488
INCREASE	411	16,808	106	305
REQUIRED	438	16,344	213	225
ASSUMED	216	4256	33	183
PRODUCES	151	2485	28	123
APPEAR	278	10754	32	246
BECOMES	246	7647	45	201
ANALYZED	85	140	3	82
DECREASE	118	1204	35	83

RESULTS AND DISCUSSION

The 30 verbs plot the graphical display in Figure 1. There are two data plots: the first plot represents the column, and the second represents the row. The column represents the corpora; each corpus is marked by a point and a label. On the other hand, the row represents all the 30 verbs from Table 1. Each is also marked by a point and a label. To have a better visualisation of the data plots, Figure 2 displays the corpora plot (column) and Figure 3 the verbs plot (row). Tables 2 and 3 provide the coordinates of both the columns (corpora) and rows (verbs) on the map respectively.

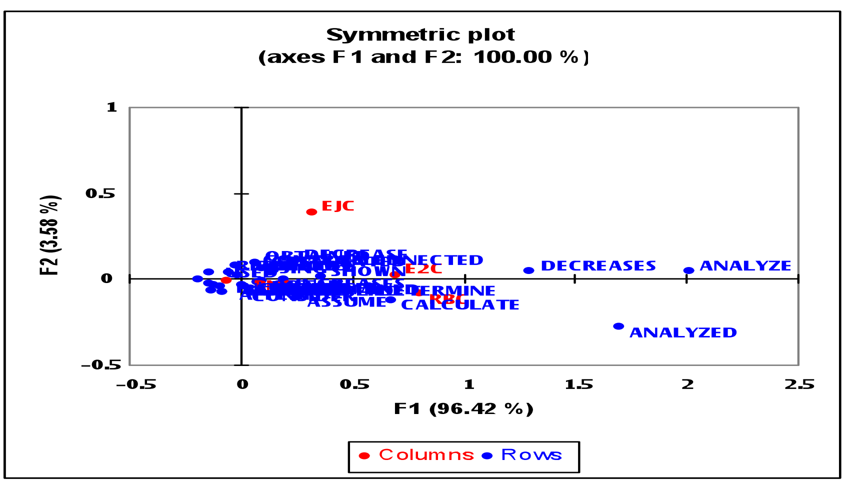

FIGURE 1. CA map of verbs (columns and rows)

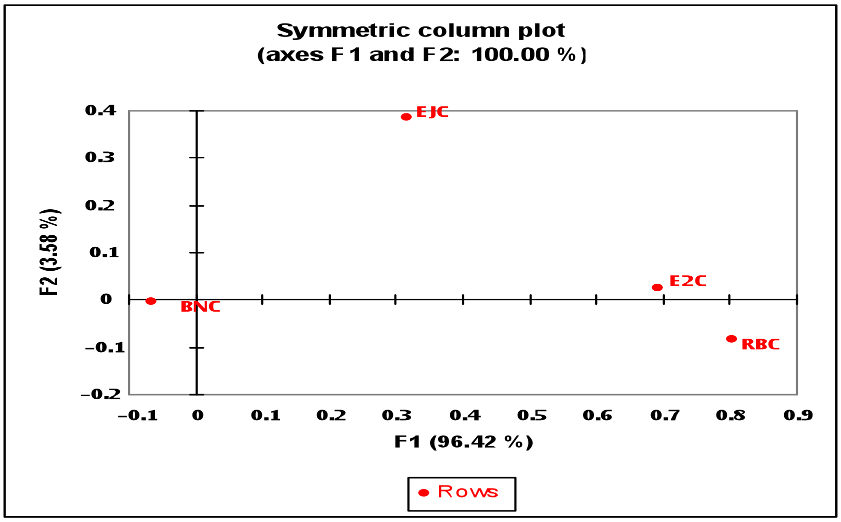

FIGURE 2. CA map of corpora (columns)

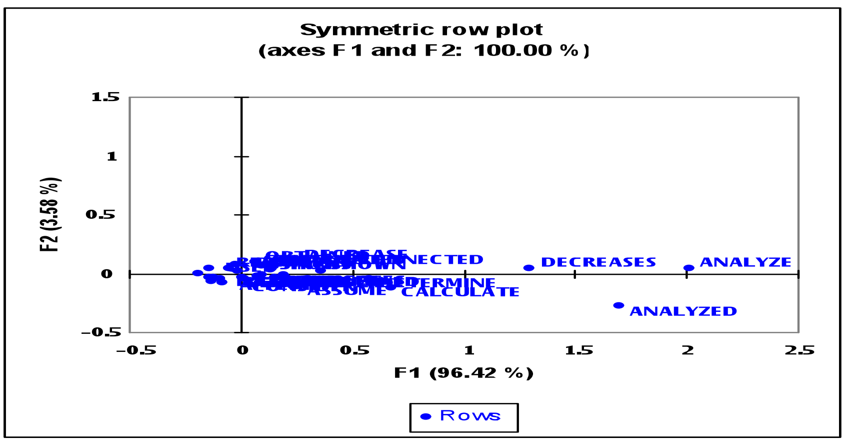

FIGURE 3. CA map of verbs (rows)

TABLE 2. Coordinates of corpora

	F1	F2
E^2C	0.691	0.027
BNC	-0.068	-0.003
EJC	0.314	0.389
RBC	0.803	-0.081

TABLE 3. Coordinates of verbs

VERBS	F1	F2
SHOWN	0.357	0.023
SHOWS	0.134	0.044
CONNECTED	0.477	0.085
DETERMINE	0.568	-0.040
USING	-0.019	0.025
APPLIED	0.296	-0.040
DEFINED	0.203	-0.025
ASSUME	0.251	-0.106
OBTAINED	0.061	0.103
DETERMINED	0.148	-0.027
CALCULATE	0.671	-0.116
NOTE	-0.002	-0.029
OBTAIN	0.086	-0.007
USED	-0.191	0.006
ANALYZE	2.007	0.050
CALCULATED	0.094	0.069
CONSIDER	-0.089	-0.069
INCREASES	0.185	0.000
PROVIDES	-0.060	0.047
CONTROLLED	-0.023	0.083
DECREASES	1.292	0.050
RESULT	-0.124	-0.031
INCREASE	-0.149	-0.022
REQUIRED	-0.148	0.045
ASSUMED	0.016	-0.037
PRODUCES	0.069	-0.013
APPEAR	-0.133	-0.059
BECOMES	-0.097	-0.036
ANALYZED	1.691	-0.270
DECREASE	0.242	0.112

The variation from left to right (along F1 axis) is 96.42% of total inertia, and the variation from top to bottom (along F2 axis) is 3.58%. Therefore, the association between the corpora and the verbs is two dimensional of a good quality at 100%. The quality of the analysis can be evaluated by consulting the table of the eigenvalues (Table 4). The eigenvalues reflect the relative importance of the dimensions; the first is always being the most important. If the sum of the two (or a few) first eigenvalues is close to the total represented, then the quality of the analysis is very high. The correspondence analysis of the 30 verbs is of good quality as the sum of the first two eigenvalues adds up to 100% of the total. The values also indicate that the differences between the corpora are mainly along the F1 axis. In other words, the main differences among the corpora can be described from the information or words along the F1 axis. The F1 axis provides the relationship between general English (BNC) and the specialised corpora, while the F2 axis presents the associations between the two academic engineering corpora: EJC and RBC.

TABLE 4. Eigenvalues and percentages of inertia

	F1	F2
Eigenvalue	0.049	0.002
Inertia (%)	96.416	3.584
Cumulative (%)	96.416	100.0

The plotted points in Figure 2 show that the verbs clearly distinguish the corpora from each other, with all the three engineering corpora in the same quadrant. RBC appears to be the most specific sub-corpus even from the use of verbs outlook (as opposed to nouns which can be categorised according to their technical meanings). On the other hand, the F2 axis reveals more on the interrelationship of EJC and RBC. E^2C (0.027) and EJC (0.389) share the same positive quadrant, whereas RBC (-0.081) in the negative quadrant; BNC, however, is in the negative quadrant for both axes (Table 2). It appears that though both academic engineering text types may share many words, the CA proves that the use of verbs in EJC and RBC still differs. Generally, Figure 3 proves that general English (BNC) is indeed different from the academic engineering texts (E^2C), and RBC from EJC. It also shows that E^2C does assimilate the features of both text types; for both axes, E^2C remains in between RBC and EJC.

The corresponding words plot in Figure 3 shows which of the 30 verbs contribute to the differences and similarities of the corpora, according to the positions reflected on both axes. Verbs which are distant from BNC contribute to the specialised corpora, such as *analyse*, *analysed*, and *decrease* with the coordinates of 2.007, 1.691 and 1.292 respectively. These are words which contribute to the differences of the specialised corpora on the F1 axis. Words which contribute to the differences on the F2 axis include *connected*, *calculate*, and *assume* with the coordinates of 0.085, -0.116 and -0.106 respectively. Words which are closer to any of the corpus contribute to that corpus; for example, the words *analyse* and *analysed* are closer to EJC. Therefore these words contribute to the genre. The further the words from BNC, the more significant they are in the specialised corpora.

The values of corpora contribution on both axes are presented in Table 5. A contribution value is the percentage of inertia (variance) of a particular dimension (or axis) which is explained by the point (Garson 2008). It shows that all the three specialised corpora do contribute to the differences on the F1 axis, with E^2C and RBC account for the most of the total information explained by the points on the F1 axis, that is 0.438 (43.8%) and 0.455 (45.5%) respectively. This implies that the use of the 30 verbs strongly characterises the engineering academic texts. However, EJC appears to have a higher contribution value on the F2 axis, 0.854 (85.4%), than the F1 axis, 0.021 (2.1%). Its opposition, the RBC, modestly contributes 12.4%. It appears again that the differences between EJC and RBC are reflected on the F2 axis.

TABLE 5. Contribution values of corpora

	F1	F2
E^2C	0.438	0.018
BNC	0.086	0.003
EJC	0.021	0.854
RBC	0.455	0.124

This implies that EJC shows its attributes more on the F2 axis. RBC, in contrast, displays more of its qualities from the F1 axis. This corresponds with earlier observation that RBC is the most specific among the other specialised corpora. The values on the F1 axis also suggest that the use of verbs in EJC (0.021) is the closest to BNC (0.086) than any other specialised corpora in the study. It means that EJC may assimilate more features of the general English than RBC, which evidently is closer to E^2C. As informed earlier, the associations of all the corpora are mainly described by the information on the axis which has a higher eigenvalue (or inertia value) among the dimensions (or axes).

The words which contribute to these similarities and differences are reflected in the word contribution values in Table 6 and Table 7. Verbs which contribute to F1 axis are listed in Table 6, and F2 axis Table 7. The higher value of a word between the axes reflects the axis to which the word contributes more. For example, in Table 6, the contribution values of the verb *used* are 0.163 on the F1 axis and 0.005 on the F2 axis; this means that the verb *used* contributes more to the analysis of the corpora on the F1 axis. Therefore, in discussing the different features between the specialised corpora and general English corpus, further linguistic analysis of the verb *used* may promise significant results.

A closer look at the verbs which contribute to the differences on the F1 axis (Table 6) reveals that *used* (16.3%) has the highest contribution value, followed by *shown* (15.3%), *determine* (10.9%), *analyse* (8.9%), *decreases* (8.1%), *connected* (6.4%), *analysed* (5.8%), *applied* (5.2%), *calculate* (4.2%), *increase* (2.6%), *defined* (1.8%), *determined* (1.2%), *increase* (1.1%), *obtain* (0.2%) and *produces* (0.1%). Nevertheless, the maps in Figures 1 and 2, and the coordinates in Table 3 provide a clearer role played by these verbs along the F1 axis. It shows that *analyse* (coordinate: 2.007), *analysed* (coordinate: 1.691) and *decreases* (coordinate: 1.292) are used distinctly in the specialised corpora, as opposed to *used* (coordinate: -0.191) and *increase* (coordinate: -0.019) which appear to make up the features of the general English more.

Similarly, on the F2 axis (Table 7), the contribution values suggest that *obtained* (13.2%) plays the most important role in differentiating EJC from RBC, followed by *consider* (10.3%), *assume* (9.5%), *appear* (6.9), *required* (6.2%), *controlled* (6.1%), *shows* (4.4%), *result* (3.8%), *provides* (3.5%), *decrease* (3.2%), *using* (2.8%), *calculated* (2.5%), *becomes* (1.8%), *note* (1.7%) and *assumed* (1.1%). On the other hand, Figures 1 and 2, and the coordinates in Table 3 throw more light on the contribution of this group of verbs. It appears that there are no verbs which characterise EJC, and the closest verb to EJC on the F2 axis is *obtained*, with the coordinate of 0.103. However, the verbs seem to cluster near E^2C (coordinate: 0.027) and RBC (coordinate: -

0.081). This suggests that while the contribution values provide the set of verbs which differentiate RBC from EJC, the identification of which verbs characterise any of the corpora can be determined by observing the CA maps (coordinates). As such, the most distinct verb in the group is *assume* (coordinate: -0.106) and *consider* (coordinate: 0.069).

TABLE 6. Verbs that contribute to F1 axis

	F1	F2
USED	0.163	0.005
SHOWN	0.153	0.017
DETERMINE	0.109	0.014
ANALYZE	0.089	0.001
DECREASES	0.081	0.003
CONNECTED	0.064	0.055
ANALYZED	0.058	0.040
APPLIED	0.052	0.025
CALCULATE	0.042	0.034
INCREASE	0.026	0.015
DEFINED	0.018	0.007
DETERMINED	0.012	0.011
INCREASES	0.011	0.000
OBTAIN	0.002	0.000
PRODUCES	0.001	0.001

TABLE 7. Verbs that contribute to F2 axis

	F1	F2
OBTAINED	0.002	0.132
CONSIDER	0.006	0.103
ASSUME	0.020	0.095
APPEAR	0.013	0.069
REQUIRED	0.025	0.062
CONTROLLED	0.000	0.061
SHOWS	0.016	0.044
RESULT	0.023	0.038
PROVIDES	0.002	0.035
DECREASE	0.006	0.032
USING	0.001	0.028
CALCULATED	0.002	0.025
BECOMES	0.005	0.018
NOTE	0.000	0.017
ASSUMED	0.000	0.011

CONCLUSION

Obviously, CA proves to be a great tool to draw empirical information regarding which word to select for the possible comparison between corpora in order to distinguish their characteristics; in addition, the analysis provides the most significant words for further linguistic description. Though there is an argument that CA is an exploratory, instead of a confirmatory technique, the task to specify appropriate variables has been done by identifying the significant words to be analysed by generating the keyword and key-keyword lists prior to the CA procedure. The generating of the significant words is done using the log likelihood analysis, which is available with the Wordsmith 6 tool used for the study. Thus, this supports the significance of the output retrieved using the CA technique in this study.

The use of the CA provides interesting insights into the complex inter-relationship between the corpora. The CA of the verbs from the E^2C key-keyword list provides a clear demarcation of all the corpora. There are distinguishable features in the use of verbs between E^2C and general English (BNC), and between RBC and EJC. The contribution values in the CA, in addition to the map displays, highlight words which are differentiated along the axes, thus, provide insights into words which give the profiles of a corpus. The analyses furnish the researcher with lists of verbs that characterise each corpus.

The application of CA is most useful for genre-based study, especially to show how the same sets of words function in different genres. The identification of words significant to a corpus allows an efficient investigation on language description to take place. The genre-based wordlists can serve as a guide for the designing and planning of more specific courses for the specialised language community at the learning institution, such as journal writing. In other words, the wordlists can provide accurate information on the language input depending on the various aims set for teaching and learning the specialised language; this is especially crucial since it is found that one genre can be more specialised than the others.

ACKNOWLEDGEMENTS

This study was sponsored by UTeM and the Ministry of Education, Malaysia.

REFERENCES

Abney, S. (1996). Statistical methods and linguistics. In J. Klavans & P. Resnik (Eds.). *The balancing act: Combining symbolic and statistical approaches to language (pp. 1-23)*. Cambridge: MIT Press.

Chujo, K., Utiyama, M., Nakamura, T. & Oghigian, K. (2010). Evaluating statistically-extracted domain-specific word lists. In G. Weir & S. Ishikawa (Eds.). *Corpus, ICT and language education (pp. 53-64)*. Glasgow: University of Strathclyde Publishing.

Fadhil H.A.Y. Mansor. (2001). A case of an extra-sensitive perception of ESP. *GEMA Online® Journal of Language Studies. Vol. 1*(1), 15-27.

Firooz Namzar, Nor Fariza Mohd Nor, Noraini Ibrahim & Jamilah Mustafa. (2012). Analysis of collocations in the Iranian postgraduate students' writings. *3L: The Southeast Asian Journal of English Language Studies. Vol. 18*(1), 11-22.

Fuentes, A. C. & Rokowski, P. E. (2003). Using corpus resources as complementary task material in ESP. *English for Specific Purposes World. Vol. 6*(2). Retrieved July 20, 2011 from http://esp-world.7p.com/articles_6/C2_.htm

Garson, G. D. (2008). Correspondence analysis. Retrieved July 31, 2009 from http://faculty.chass.ncsu.edu/garson/PA765/correspondence.htm

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

Harwood, N. (2002). Taking a lexical approach to teaching: Principles and problems. *International Journal of Applied Linguistics. Vol. 12*(2), 139-155.

Imran Ho-Abdullah. (2009). Pemantapan dan pembinaan ilmu linguistik berasaskan korpus: transformasi statistik senarai kekerapan kata. Paper presented at SKALI Seminar. Bangi: UKM, Mac.

Imran Ho-Abdullah & Laman, C. (1997). Comparing word frequencies across corpora: a correspondence analysis of varieties of English. Paper presented at the 4th New Zealand National Postgraduate Conference. Dunedin, November.

Jin, N.Y., Tong, C. S., Mariam Mohamed Nor, Mohd Ariff Ahmad Tarmizi & Alif Fairus Nor Mohamad. (2012). Corpus based analysis of the TOEFL course books: What are the words we should teach our students?. *International Review of Social Sciences and Humanities. Vol. 3*(2), 152-160.

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.

Lee, D. Y. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology. Vol. 5*(3), 37-72.

Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward.* England: Language Teaching Publication.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics. Vol. 33*(3), 299-320.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

Menon, S. & Mukundan, J. (2012). Collocations of high frequency noun keywords in Prescribed science textbooks. *International Education Studies*. *Vol. 5*(6), 149-160.

Monti, J., Barreiro, A., Elia, A., Marano, F., & Napoli, A. (2011). Taking on new challenges in multi-word unit processing for machine translation. Paper presented at the Second International Workshop on Free/Open-Source Rule-Based Machine Translation. Retrieved January 23, 2014 from http://hdl. handle. net/10609/5646

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

Nattinger, J. & DeCarrico, J. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

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

Nishina, Y. (2007). A Corpus-driven approach to genre analysis: The reinvestigation of academic, newspaper and literary texts. *Empirical Language Research (ELR) Journal. Vol. 2*(1). Retrieved July 31, 2008 from http://ejournals.org.uk/ELR/article/2007/2

Noorli Khamis & Imran Ho-Abdullah. (2012). Correspondence analysis: Comparing wordlists across specialised corpora. Paper presented at CHUSER 2012. Malaysia: Kota Kinabalu, December.

Nurul Farahin Musa & Noorli Khamis. (2014). Features of engineering research articles. *Science International (Special Issue). Vol. 26*(4), 1557-1561.

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

Smith, C., Butler, N. L., Griffith, K. G. & Kritsonis, W. A. (2007). The role of communication context, corpus-based grammar, and scaffolded interaction in ESL/EFL instruction. *The Lamar University Electronic Journal of Student Research 4*. Retrieved September 9, 2008 from http://www.eric.ed.gov/PDFS/ED495290.pdf

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.