# Lexical Features of Engineering English vs. General English

*Noorli Khamis*
noorli@utem.edu.my
*Centre for Languages and Human Development (PBPI)*
*Universiti Teknikal Malaysia Melaka (UTeM)*

*Imran Ho-Abdullah*
imranho@pkrisc.cc.ukm.my
*School of Language Studies and Linguistics*
*Universiti Kebangsaan Malaysia*

## ABSTRACT

The knowledge on the features of the English varieties is essential to understand the differences and similarities of the varieties for second language teaching and learning, either for general proficiency (EGP) or English for Specific Purposes (ESP) classes. This paper demonstrates a corpus-based comparison of the lexical features between an ESP variety (Engineering English) and a General English (GE). Two corpora are used in the study; the Engineering English Corpus (EEC) acts as the representation of the specialized language, and the British National Corpus (BNC) as the General English (GE). The analyses are conducted by employing the *WordList* functions of a linguistic software – Wordsmith. Discussions on the differences (or similarities) of these two corpora include general statistics, text coverage and vocabulary size. The empirical findings in this study highlight the general lexical features of both corpora. The analyses verify that the Engineering English has less varied vocabulary, but higher text coverage than the GE; in other words, most of the words are used repeatedly throughout the EEC. Thus, this study further emphasizes the importance of corpus-based lexical investigations in providing empirical evidences for language description.

**Keywords:** corpus; lexical features; specialized corpus; language description; ESP

## INTRODUCTION

The study of English for Specific Purposes (ESP) language features, in particular the lexical features, facilitates ESP teaching and learning. Peters and Fernández (2013) asserts that ESP is particularly different from General English (GE) or English for General Purposes (EGP), because ESP deals with lesser or limited number of varieties, text types and situations; in most cases, it will be one at a time. Bowker and Pearson (2002) claim that apart from the obvious attribute of an ESP, i.e. specialized vocabulary, another salient feature of an ESP is the distinctive ways of combining words and arranging information that are different from GE. This includes collocations and stylistic features (Sadeghi & Nobakht, 2014).

In corpus-based studies, it has been established that there are words used repeatedly throughout a text, and these words, therefore, have high frequencies in a language. If a very large proportion of these high frequency words in a text can be identified, it allows a good degree of comprehension when reading a text. If that is the case, the common questions asked include how many words do a student need to know to be able to comprehend a text?, and do some disciplines use a greater range of different words?. These and other questions related to descriptions of word use in a language lead to answers, which are essential to ensure the effectiveness of second language teaching and learning, for either general proficiency (EGP) or English for Specific Purposes (ESP) classes.

eISSN: 2550-2131
ISSN: 1675-8021

This study adopts a corpus-based comparison of the lexical features between an ESP variety (an Engineering English) by comparing it with the GE, which is represented by the British National Corpus (BNC). The findings demonstrate the general lexical features of both, which provide insights into differences, between an ESP and a GE corpus empirically. The discussion highlights the importance of knowledge on lexical features in assisting ESP course designers, material developers and language instructors to make more informed decisions on, for example, selection of contents, construction of syllabus, preparation of teaching materials, and preparation of assignments and exam questions

## CORPUS LINGUISTICS AND ESP

### GENERAL DESCRIPTIVE CORPORA VS. SPECIALIZED CORPORA

General descriptive corpora are designed for various investigations into linguistic features – lexis, grammar, discourse pattern, pragmatics or prosody of a language (Wilkinson, 2014). The construction of a general descriptive corpus is usually determined by sampling criteria to gain reliable representation of a target language. The sampling criterion can be observed from a collection of texts from various genres and text types, for example, the construction of the BNC, which involves the collections of written and spoken British English from newspapers, specialist periodicals, journals, academic books, fictions, published and unpublished letters, memoranda, school and university essays, as well as unscripted informal conversations of volunteers representing various ages, regions and social classes. The contexts range from formal to informal situations. The criteria set to construct this type of corpus are obviously to fulfil the aim of representing "... a whole language or a geographical variety" (Gavioli, 2005: 7).

On the other hand, the specialized corpora represent specialized text-types and topics, which are created for a specific teaching and/or learning interest. Unlike the general descriptive corpora, the specialized 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 content 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 specialized categories such as newspaper language or academic language

In other words, a specialized corpus is designed for a sample collection of a sub-language, such as the collection of research articles on a single topic *Semiconductor Diodes* by different authors. Specialized corpora can also involve a collection of texts for academic language description – specialized academic corpora, which may take into consideration written or spoken language, published materials, gender of authors, etc. depending on the objective of the corpus construction (Gavioli, 2005). Such a corpus is the British Academic Writing English (BAWE), which comprises students' writing in British education, and it is evenly distributed across 35 disciplines. Both corpora are specialized academic language, however, the former is more specialized and 'restricted' than the latter, because it is a collection of texts not only of a sub-language, but also on a single topic.

Due to this nature, the specialized corpora are not valid to be used to carry out observations on general use of language. Nevertheless, they are commonly created for conducting comparative investigation on language features between specialized language and general language (Bowker & Pearson, 2002). In fact, specialized corpora are regarded as more

influential and accurate to describe features of specialized language than general descriptive corpora (Noorli Khamis & Imran Ho-Abdullah, 2015).

## SPECIALIZED CORPORA FOR ESP LANGUAGE INVESTIGATIONS

English for Specific Purposes (ESP) is defined as the study of the English language in specialized contexts and fields of knowledge, such as medicine, engineering, business and the like (Triki, 2002) . The descriptions of ESP may include grammar, lexis, register, study skills, discourse and genre. This can be taken as the study of English language as it is used in a particular domain; as such, it may display some distinctive usage from GE.

Bowker and Pearson (2002) define ESP as the language that is used to discuss specialized fields of knowledge. They maintain that there is some degree of overlap between GE and ESP due to the fact that a lot of GE words would also appear in a specific domain usage. In spite of this, there may also exist "... special ways of combining terms or of arranging information" (2002, p. 26) that are different from GE. A language for specific purposes has a set of specialized terms, which may be combined in a distinctive way to construct meaning in the domain. Some of these combinations are called *specialized vocabulary*. Bowker and Pearson also highlight that the purpose of ESP is to aid discussions among speakers in a specialized field of knowledge.

Gavioli (2005) regards ESP and specialized corpora as one happy marriage. He emphasises the relevance of quantitative data in not only general English language description, but even more so in ESP. The quantitative attribute in ESP language description is also highlighted by Halliday, who considers 'scientific English' as a generalised functional variety, or register. "... a register is a cluster of associated features having a greater-than-random (or rather than predicted by their unconditioned probabilities) tendency to co-occur" (1996, p. 54).

With this quantitative quality, a corpus-based investigation into the varieties of ESP is regarded as a practical and interesting approach. It is because in ESP courses, identifying 'what to teach' seems to be a central issue amongst the language instructors, especially due to the fact that most ESP varieties contain different teaching environments or genres to cater to. In other words, ESP has very wide ranges of teaching situations and learners' needs. McEnery and Wilson (2001, p. 121) view ESP as a particular domain-specific area of language teaching and learning, where "... corpora can be used to provide many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain more directly than those taken from more general language corpora".

Corpus work allows the focus of study on a more restricted topic range and text types. Therefore, specialized corpora have become a more reliable tool to describe the specific language and its sublanguages. This suggests the potential of corpora to solve the ESP problems, which traditionally have been the concern of ESP instructors. The data from a specialized corpus can be used to determine any recurrent patterns which characterise the specialized language. More accurate explanations for the uses of lexis and phraseologies, or particular structures can be generated in order to aid language instructors in dealing with language descriptions where grammar or dictionary explanation is unable to be applied.

Partington (1998), on the same note, adds that specialized corpora are more relevant for studies on language teaching than larger general corpora because:

a. word and structure frequencies of a specialized corpus are greatly different from a large corpus of General English; for example, an Encarta micro-corpus of health articles, with only 24,805 words, provides 33 samples of the word

*cancer* in context. In contrast, the BNC Sampler written components, with 1,000,000 words, contains no sample of the word *cancer* at all (Tribble, 1997).

b. some common language functions, which are typical in specialized language texts, can be dealt with by observing the patterns identified; for example, the use of multi-word units (of nominal groups) in a technical corpus of English (Thouvenin, 1996)
c. the study of particular genre of a specialized language, like books or research articles, can highlight the features to be acquired by the language learners
d. problematic lexis or phraseologies for non-native speakers can be examined to gain insights into useful information on their uses and functions in the specialized language; for example, sub-technical words that are used to describe subject-specific concepts in a text.

Hunston (2000) strongly recommends the use of corpus-based investigations for ESP language description by claiming that corpus study offers different, yet significant contributions in its methodology than any other linguistics studies. Among the qualities discussed are:

a. the use of authentic data in generating the patterns used in the specialized language
b. the collection of data (or texts), which come from wide ranges of sources reflects the real language use, as opposed to selected data collection from linguistics grounds, which reflects the researcher's aim of study
c. the collection of abundance of data provides detailed samples of language use according to contexts and senses
d. the systematic organisation of the data allows statistical descriptions to be carried out and replicated for further research purposes
e. the data are analysed according its natural occurrence instead of the available linguistic theories.

As such, employing the corpus-based approach for a specialized language description is seen as a preferred option.

## THIS STUDY

This study demonstrates a corpus-based comparison of the general lexical features of an ESP variety (Engineering English) and a GE. The description for this study involves 2 aspects, i.e. the vocabulary size and text coverage of both corpora. Vocabulary size refers to the number of vocabulary that a reader needs to have to be able to read a text adequately, while lexical coverage is the percentage of known words in the text.

It has been established that vocabulary range can determine the reading proficiency of a language learner (Laufer & Ravenhorst-Kalovski, 2010; Sen & Kuleli, 2015; Teng, 2016). Many studies have looked into the issues of vocabulary size and text coverage in describing the lexical needs of language learners. Waring and Nation (1997), for example, discuss the estimates of the vocabulary size and their significance for second language learners by providing a thorough survey of previous research on the issue. Several thresholds have been suggested, which include, the need to have around 20,000 word families for university graduate, 3,000 or so of high frequency words in a language for undergraduates, and 2,000 for EAP students (Schmitt & Schmitt, 2014). A study by Hsueh-chao and Nation (2000) provides an experimental support that learners need to have around 98% of words coverage in a text to be able to read with ease. Earlier, Na and Nation (1985) had set the

threshold at 95% for a reasonable comprehension of a text. However, the vocabulary size of 2,000 words is the best selection for English learners to memorize (Nation & Kyongho, 1995). Of course, this does not mean that vocabulary knowledge is the sole skill to gain adequate comprehension, but it does indicate that vocabulary knowledge is a critical component in language skills, such as reading, writing, speaking and listening.

## METHOD

There are two corpora, which reflected the specialized (ESP) language and General English in this study. The specialized corpus is created by the researcher, while the GE corpus is obtained online. The corpora were:

a) the Engineering English Corpus (EEC)
b) British National Corpus (BNC)

EEC serves as a specialized corpus for comparative language investigation with the BNC, as the GE.

### ENGINEERING ENGLISH CORPUS (EEC)

#### PRINCIPLES FOR THE SELECTION OF TEXT TYPES

The texts selected for this corpus consisted of two genres of an Engineering English variety, i.e. the Electronics and Computer Engineering English. It is referred as the Engineering English henceforth. The two genres of written texts are **suggested reference books** for the course (Electronics and Computer Engineering) and **e-journals articles**. These two genres of written texts were chosen with the following rationales:

a) The language investigation carried out with these two written text types involves a social purpose, that is, to provide a general lexical profiling of an Engineering English variety. The analysis of these text types is conducted with provide insights into the teaching and learning of an Engineering English.
b) The two genres were selected due to the fact that these text types are the written materials at the disposal of the learners. The suggested reference books and e-journals are regarded as a representative collection of the subject areas in the learning context of the learners (Tribble, 1997); thus, they provide the right corpus to answer questions on technical and academic lexis of the Engineering English. In addition, the fact that the reading materials (textbooks) are suggested to the learners proves that the texts are "... actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences" (Stubbs, 1996). The selection of the texts for e-journals was also based on the topics provided in the suggested reading materials.
c) Full texts were collected from the two genres in order to capture as many linguistic features of the Engineering English as possible to be analysed textbooks (Scott & Tribble, 2006; Meyer, 2002; Bowker & Pearson, 2002).
d) In this study, the British National Corpus (BNC) is used to represent the General English. The comparison between EEC and GE provides insights into the lexical features which constitute the characteristics of the Engineering English.

eISSN: 2550-2131
ISSN: 1675-8021

## THE CORPUS

This corpus acts as the representation of the Engineering English for the study. It consists of 102 texts with the size of running words at 677,993.

The reference books for this corpus were identified from the handbook of an Engineering faculty in one of the local universities. The handbook contained suggested reference books for the students from all the programs 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 programs. 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 library. There are 34 texts, with 425,854 running words, in this corpus. These texts are actually the total of chapters from both textbooks.

Another composition of this corpus is the engineering journal articles, collected from the online databases of the same local university. The links selected for this study are:

a) ASME Online journals
b) ScienceDirect
c) IEEE Xplore
d) Wilson Applied Science & Technology

These online databases provide full articles for engineering articles. The journals are selected based on the titles of chapters in the reference books. With 252,139 running words, there are 68 journal articles selected for this corpus. The search for articles from the online databases was conducted by keying-in the key words (content words, without function words such as *the*, *and*, *in*) 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. The distribution of articles according to the databases from which they were retrieved is as in Table 1. All the articles are of full-text version. The length of the selected articles ranges from four to seven pages. Due to limited subscription, the publication period are kept from 1995 onwards. Table 2 provides the final composition of the EEC.

TABLE 1. The distribution of retrieved journal articles

| Databases | Book 1 | Book 2 | Total |
|---|---|---|---|
| ACME Online journals | 9 | 8 | 17 |
| ScienceDirect | 8 | 9 | 17 |
| IEEE Xplore | 9 | 8 | 17 |
| Wilson Applied Science & Technology | 8 | 9 | 17 |
| Total | 34 | 34 | 68 |

TABLE 2. The composition of EEC

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

## BRITISH NATIONAL CORPUS (BNC)

This corpus consists of 100 million running words, 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, journals, academic books, fictions, published and

unpublished letters, memoranda, and, school and university essays. 10% of the corpus, which comprises the spoken samples, is taken from unscripted informal conversations of volunteers representing various ages, regions and social classes. Apart from that, the samples are also collected from other different contexts, including formal situations, like business and government meetings, to informal situations, like radio shows. Also, BNC is a modern mega-corpus, which licence is easily obtained online at http://bncweb.info/.

In this study, the BNC acts as a reference corpus to obtain any statistical information on the spread of the lexical behaviours exist in the specialized corpus, thus, proving whether the identified features 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 EEC.

### DATA ANALYSIS SOFTWARE - WORDSMITH

The *Wordsmith* software is a multi-function software package, which offers programs for investigating the lexical behaviour in either a single text or a large corpus. This software provides the point of departure for the whole investigation. Developed by Scott (2006), it is released by Oxford University Press (OUP). 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). Thus, it has been employed by many corpus linguists for their study. It is an excellent software, and also available on the Internet.

This software features the wordlist, keyword, and concordance programs. These programs are used by the OUP for their own work in lexicography for preparing dictionaries. These three main programs, wordlists, keyword, and concordance, offer various interesting and remarkable features that are useful for language investigation.

However, this study mainly employ the functions of the *WordList* program to provide useful details, which include the running words (tokens), types (distinct words), STTR (standardised type token ratio), mean word length, *n*-letter words etc. This statistical information offers the basic lexical features of the corpus for investigation.

## DISCUSSION

The analysis begins with the general statistics of the corpora. Next, the discussions on the aspects, i.e. text coverage and vocabulary size, will be provided as an attempt to provide a general lexical comparison between the corpora.

### GENERAL STATISTICS

The statistical details of a corpus should be discussed before further analysis takes place. The basic statistical data of the corpora are retrieved from the *Wordlist* program. These sets of information are useful to provide initial insights into understanding the relative variety of vocabulary and the general differences of the corpora. The details are as in Table 3.

TABLE 3. Basic statistical data of BNC and EEC

| Statistical Details | BNC | EEC |
|---|---|---|
| tokens used for word list | 97,860,872 | 601,481 |
| types (distinct words) | 512,588 | 12,458 |
| standardised TTR | 43 | 30 |
| mean word length (in characters) | 5 | 4.85 |
| Ratio of 1-4 letter words | 58% | 54.16% |

To have a meaningful interpretation of the results, the statistical information of BNC, as the reference corpus, is compared with the information of EEC. The comparison between both corpora is relevant to investigate any possible similarities or differences between EEC and GE. Though many other studies specified the use of written BNC as reference corpus in making comparisons with other written corpora, this study did not attempt to make such distinction because the concern is to examine any possible discrepancies or similarities between EEC with the General English. In other words, the issue whether there are differences or similarities between EEC with the spoken or written General English, is not discussed in this study.

As shown in Table 3, EEC consists of 102 texts with a total of 601,481 words or tokens, and 12,458 different words or types. In comparison, BNC consists of 97,860,872 tokens and 512,588 word types. A valid comparison can be observed from the standardised type/token ratio (Standardised TTR or STTR) values of both corpora. The STTR is obtained by computing the type/token ratio for the first 1000 words in the corpus, and for the following sets of 1000 words to the end of the corpus. A running average is computed, and the standardised type/token ratio is obtained. STTR is an interesting measure because with it, comparing corpora of differing lengths is possible; it segments the corpus into comparable chunks and calculates the type/token ratio for each. Thus, STTR is reliable to be observed as part of this investigation.

STTR suggests the lexical variation or diversity of the corpus. 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. There are about 30 distinct words used in every 1000 tokens for EEC. In contrast, BNC has 13 more different words for every 1000 words (43 types). This initial statistics information suggests that EEC has less varied vocabulary than BNC in every 1000 words. This may be accounted by the characteristics of EEC as a specific domain corpus; the specific areas or topics result in more specific and lesser words to be used.

An interesting finding is noted in the next statistics information, the mean word-length. Despite the different views researchers maintain of this statistics information, Nishina (2007) asserts that word-length can be a useful index to investigate text difficulty and stylistics. The higher the value of the average word length, the more difficult the readability of the text. The use of longer words is taken to mean that the target texts have many difficult words from a solely empirical perspective. However, from Table 2, it was found that EEC has almost the same word-length average as BNC, that is 4.85 (EEC) to 5 (BNC) characters. This suggests that generally, EEC has the same level of readability as BNC from the empirical point of view. EEC is generally not made up of longer words, which suggests the same difficulty or complexity level of its texts with any other General English texts.

This same notion is also suggested by the ratio of 1-4 letter words, which reveals a relatively small difference in both corpora (54.16% for EEC and 58% for BNC). A lower value of the ratio of 1-4 letter words represents a more difficult text. Therefore, the ratio values imply that the difficulty level of EEC is quite similar to GE. The small difference (3.84%) suggests that EEC could be slightly difficult than GE; this can be accounted by the use of its technical and/or sub technical words.

## TEXT COVERAGE

Figure 1 provides the text coverage of the first 2000 word types in both corpora. With 2000 word types, EEC has a text coverage of 92%, in comparison with BNC, which has approximately a 76% of text coverage. This is not surprising because EEC is a specialized corpus, and therefore a lesser number of words should be able to reasonably cover a good range of a text. This finding corresponds with other ESP corpora investigations such as

conducted by Someya (1999), with Business English, and Farell (1990) with Electronics English. It is concluded that in each ESP text, its own specialized topic vocabulary is used at a very high rate of frequency to convey the unique message of the text. However, it should be noted that, those studies were done with lemmatised word types. Even so, this study reveals that whatever the benchmark is, the general frequency curve does not decline at a regular rate across the corpus. It still reflects the same notion that the high frequency words cover the massive bulk of a corpus. Figure 2 plots the coverage achieved by word forms with every 2000th words.

Figure 2 shows not a different frequency curve from other language studies. The rapid frequency decline in corpora has been accounted by the famous Zipf's Law. The Zipf's Law model predicts a very rapid decrease in frequency among the most frequent words, which become slower as the rank grows, leaving very long tail of words with similar low frequencies, with roughly half occurring once only as hapax legomena (words occurring only once in a text). This, too, applies to EEC; the distribution obeys a 'power law' (Scott & Tribble, 2006). The two underlying principles of the law are the 'Force of Unification' (an economy of effort, i.e. speakers extremely often opt for well-known high-frequency words) and the 'Force of Diversification' (the need for distinct words). The 'Force of Unification' accounts for the high frequency words, and the 'Force of Diversification' for the low frequency words and, for certain specialized corpora, hapax legomena.

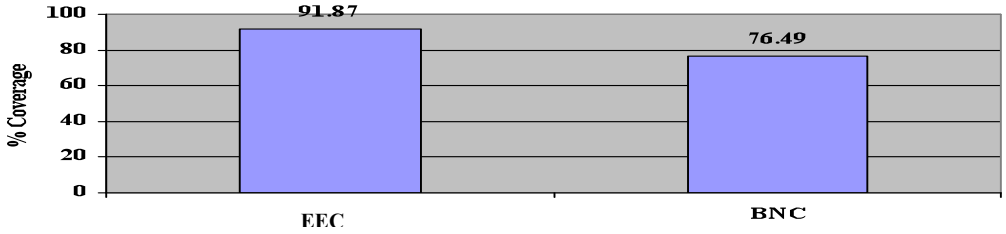


FIGURE 1. Text coverage of the first 2000 frequent word types in EEC and BNC

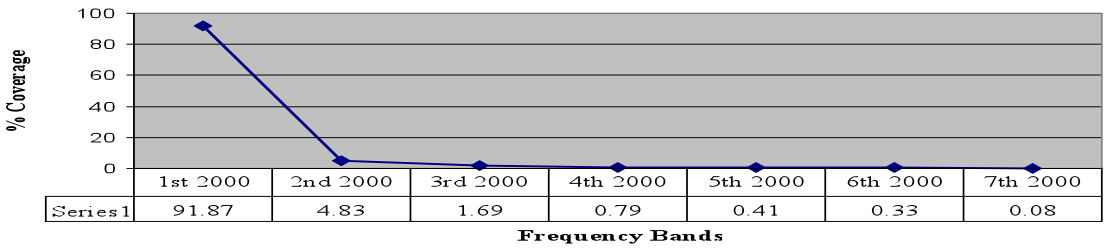


FIGURE 2. Text coverage of every 2000th word types in EEC

## VOCABULARY SIZE

Another investigation that can be done to discover the general description of EEC is its vocabulary size. This is also carried out by comparing EEC with BNC. However, Biber (2006) cautions that in comparing vocabulary distribution across corpora of different sizes, the problem is that there can be a misleading information since smaller corpora seem to use a larger stock of different words than larger corpora, because words tend to be repeated in larger corpora. In other words, the word type distributions are not linear relationships. Therefore, to compensate for this difference, all raw frequency counts need to be 'normalised' to a rate of occurrence per 1 million words. These normalised rates can then be compared directly across corpora. The choice of norming to a specific figure, for example 1,000, 10,000 or 100,000 words, is arbitrary. However, according to Meyer (2002), as larger numbers and corpora are analysed, norming to a higher figure is encouraged. For this study,

the frequency counts were normalised at the occurrences per 1,000,000. The normalised frequency for the raw data was obtained with the following formula:

**Normalised frequency = (absolute frequency/corpus size) X 1,000,000**

The formula proposed by Biber (2006) to estimate the normalised number of word types in a corpus is:

**Normalised # of word types = (# of word types / square root of corpus size) X 1,000**

Therefore, the normalised word type values for all the corpora in this study are as in Table 4.

TABLE 4. Normalised values of word types for BNC and EEC

| | BNC | EEC |
|---|---|---|
| tokens used for word list | 97,860,872 | 601,481 |
| types (distinct words) | 512,588 | 12,458 |
| Normed # of word types (per million words) | **51,816** | **16,063** |

The normalised procedure is used to compare patterns of word use between corpora. This rule is only a general approximation, and it may vary for extremely small or large corpus. Nevertheless, it can be used for good estimation for comparison across moderate-sized corpora. Biber (2006) provides detailed experiments on the effects of corpus size on the apparent vocabulary diversity. Also, because the norming of word type counts provides only the approximate value of the non-linear relationship, Biber posits that the detailed analysis of individual words would be inappropriate. Nevertheless, major trends of the lexical profiles such as in this section henceforth can be captured by this procedure.

Figure 3 compares the vocabulary growth curves between EEC and BNC. Although the numbers of word types differ greatly, the growth curves of both corpora show a relatively similar pattern, in which they grow slowly at the beginning of the graph, before going steady and rising significantly after 700,000 word tokens for BNC and 800,000 word tokens for EEC. At these points, the numbers of word types are 970 for BNC and 670 for EEC. Table 5 provides the type-token comparison in figures. This graph also reveals that both corpora do not show any sign of significant slowdown in their vocabulary growth. The observed pattern suggests that new words are most likely to be added to the lists. In other words, EEC is 'lexically opened' (Someya, 1999). This is very true for EEC because of its corpus design and size. If more suggested engineering textbooks and online journals, or other sources from other genres, are added to the corpus, there is a high chance more word types will be discovered. This finding is relevant because for other ESP, for example Business Letter Corpus (BLC), Someya finds that the growth of the corpus was minimal after reaching half a million of word tokens, suggesting that the ESP variety was much more 'lexically closed' - a very restricted number of vocabulary, than a general English (the comparison was done with Brown and LOB Corpus as reference corpora).

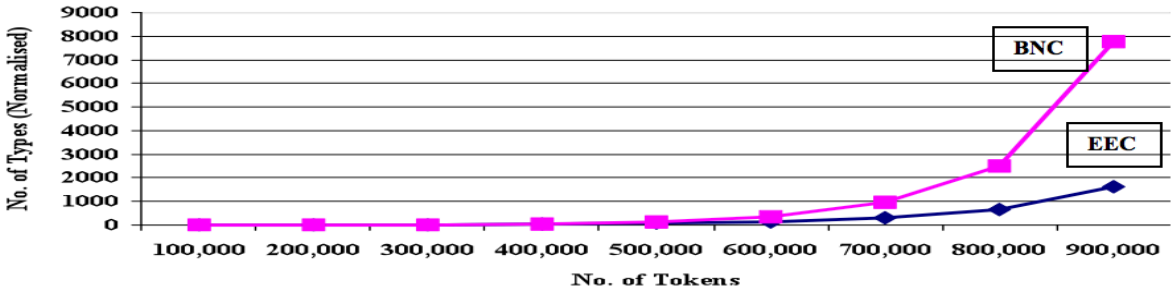


FIGURE 3. Comparison of vocabulary growth curves

TABLE 5. Type-token comparison of EEC and BNC (normalised)

| Number of Word Tokens | Cumulative Number of Word Types | | | |
|---|---|---|---|---|
| | EEC | Type/Token (%) | BNC | Type/Token (%) |
| **100,000** | 2 | 0.002 | 2 | 0.002 |
| **200,000** | 5 | 0.003 | 7 | 0.004 |
| **300,000** | 11 | 0.004 | 19 | 0.006 |
| **400,000** | 28 | 0.007 | 47 | 0.012 |
| **500,000** | 70 | 0.014 | 124 | 0.025 |
| **600,000** | 152 | 0.025 | 365 | 0.061 |
| **700,000** | 313 | 0.045 | 970 | 0.139 |
| **800,000** | 670 | 0.084 | 2,483 | 0.310 |
| **900,000** | 1,630 | 0.181 | 7,749 | 0.861 |
| **1,000,000** | 12,458 | 1.246 | 512,588 | 512.588 |

Despite the fact that EEC is 'lexically opened', a trait shared with BNC/GE, the size of vocabulary required to reach a certain percentage of text coverage is relatively smaller in EEC than in GE. Figure 4 illustrates the relationship between the numbers of word types to the total word token for EEC and BNC. The first 1,000 word types cover approximately 85% of EEC, and about 70% of BNC. Table 6 shows that to reach the text coverage of 95%, about 3,000 word types should be attained in EEC. This differs with the accepted number of 2,000 words for English learners as discussed earlier. This difference may be due to the size of the words in the corpus. An observation by Chujo and Utimaya (2005) revealed that there is a high possibility that text coverage is more stable when vocabulary size is larger, text length is longer, and more samples (from more genres in the target language) are taken.

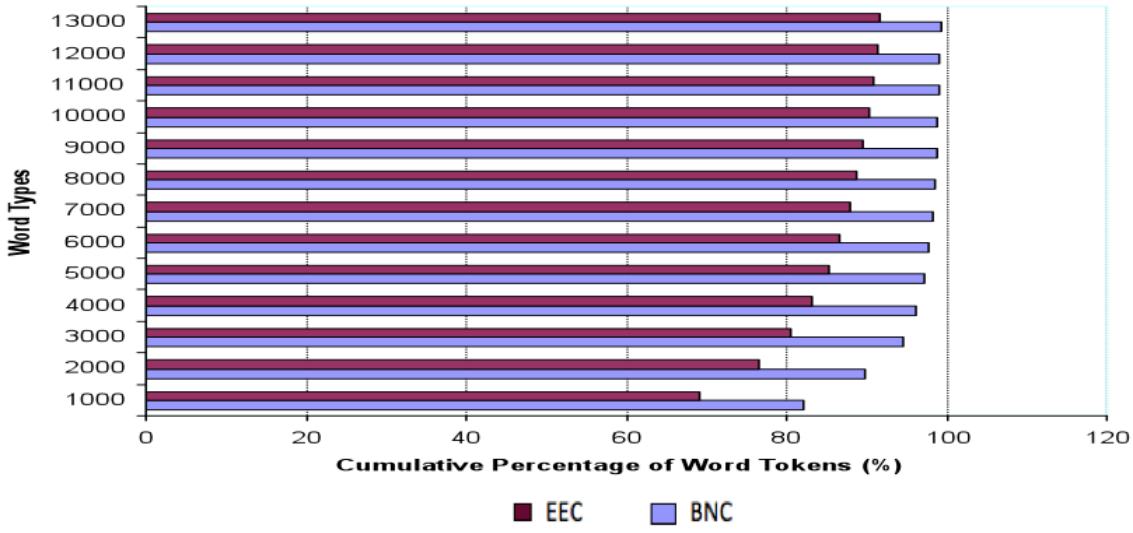


FIGURE 4. Comparison of the cumulative percentage of word tokens and total of word types

TABLE 6. Cumulative percentage of word tokens with every 500th word types

| No. of Word Types | EEC | |
|---|---|---|
| | Word Tokens (601,481 tokens) | Cumulative Percentage |
| 500 | 458,318 | 76.20 |
| 1000 | 510,284 | 84.84 |
| 1500 | 536,436 | 89.19 |
| 2000 | 552,571 | 91.87 |

| | | |
|---|---|---|
| 2500 | 563,436 | 93.67 |
| **3000** | **571,322** | **94.99** |
| 3500 | 577,165 | 95.96 |
| 4000 | 581,611 | 96.70 |
| 4500 | 585,009 | 97.26 |
| 5000 | 587,749 | 97.72 |

## CONCLUSION

Though at the initial stage it is found that EEC has less varied vocabulary than GE, the general statistical details suggest that the specialized corpus may have the same level of readability which, in turn, suggests the same level of difficulty or complexity with GE. It is proposed that the specific areas or topics in EEC result in more specific and lesser words to be used. This empirical observation implies that there is a high chance that learners could learn the specialized language the way they learn General English. Thus, this paper underlines the importance of lexis for language description.

Other salient findings that make up the profiles of EEC in comparison with GE are summarised as follows:

a) EEC has less varied vocabulary, but higher text coverage. It also indicates that most of the words are repeatedly used throughout the corpus.
b) Though the accepted threshold of text coverage and words for English learners is set at 95% and 2,000 words, EEC shows more words are definitely needed to reach such coverage in its texts. However, GE, apparently needs more words than EEC for a 95% coverage.
c) The similar vocabulary growth with GE suggests that EEC is lexically opened - a feature a few ESP varieties do not have.

These identified features entail further examinations to be made on the lexical profiles of EEC. This study also underlines the needs to conduct similar lexical investigations on other specialized corpora. The next level of lexical investigation should focus on the vocabulary types and examples of the specialised corpus for more specific and detailed descriptions of its lexical behaviours; thus, more meaningful information can be made for ESP classrooms.

## ACKNOWLEDGEMENT

This paper is funded by the university grant PJP/2017/PBPI-CTED/S01514. We would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for the support.

## REFERENCES

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers.* Amsterdam: John Benjamins Publishing Company.

Bowker, L. & Pearson, J. (2002). *Working with Specialised Language: A Practical Guide to Using Corpora*. London: Routledge.

Chujo, K. & Utiyama, M. (2005). Understanding the Role of Text Length, Sample Size and Vocabulary Size in Determining Text Coverage. *Reading in a Foreign Language. Vol. 17*(1), 1-22.

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

Hsueh-chao, M. H. & P. Nation. (2000). Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language. Vol. 13*(1), 403-430.

Hunston, S. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English.* Philadelphia: John Benjamins Publishing Co.

Laufer, B. & Ravenhorst-Kalovski, G. C. (2010). Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language. Vol. 22*(1), 15-30. Retrieved December 27, 2016 from http://files.eric.ed.gov/fulltext/EJ887873.pdf

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

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

Na, L. & Nation, P. (1985). Factors Affecting Guessing Vocabulary in Context. *RELC Journal. Vol. 16*, 33-42.

Nation, P. & Kyongho, H. (1995). Where Would General Service Vocabulary Stop and Special Purposes Vocabulary Begin?. *System. Vol. 23*(1), 35-41.

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, 2010 from http://ejournals.org.uk/ELR/article/2007/2

Noorli Khamis. & Imran Ho-Abdullah. (2015). Exploring Word Associations in Academic Engineering Texts. *3L: Language Linguistics Literature®, Southeast Asian Journal of English Language Studies. Vol. 21*(1), 117-131.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins Publishing Co.

Peters, P. & Fernández, T. (2013). The Lexical Needs of ESP Students in a Professional Field. *English for Specific Purposes. Vol. 32*, 236-247. Retrieved June 31, 2016 from http://ac.els-cdn.com/S0889490613000355/1-s2.0-S0889490613000355-main.pdf?_tid=9b1403fe-0add-11e7-b0de-00000aab0f01&acdnat=1489733436_94408bc049d9ad1699ca7d9fee 376024

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.

Sardinha, B. (1996). Review: Wordsmith Tools. *Computers & Texts 12*. Retrieved July 25, 2011 from http://users.ox.ac.uk/~ctitext2/publish/comtxt/ct12/sardinha.html

Schmitt, N. &, Schmitt, D. (2014). A Reassessment of Frequency and Vocabulary Size in L2 Vocabulary Teaching1. *Language Teaching. Vol. 47*(4), 484–503. Retrieved November 17, 2016 from https://doi. org/10.1017/S0261444812000018

Scott, M. & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Amsterdam: John Benjamins Publishing Company.

Scott, M. (2006). *Oxford Wordsmith Tools Version 4.0 Manual.* Oxford: Oxford University Press.

Şen, Y. & Kuleli, M. (2015). The Effect of Vocabulary Size and Vocabulary Depth on Reading in EFL Context. *Procedia - Social and Behavioral Sciences. Vol. 199*, 555-562. Retrieved May 7, 2016 from https://doi.org/10.1016/j.sbspro.2015.07.546

Someya, Y. (1999). A corpus-based study of lexical and grammatical features of written business English. Unpublished M.A. thesis, University of Tokyo.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*. Oxford: Blackwell Publishers.

Teng, F. (2016). The Effects of Word Exposure Frequency on Incidental Learning of the Depth of Vocabulary Knowledge. *GEMA Online® Journal of Language Studies. Vol. 16*(3), 53-70.

Thouvenin, S. P. (1996). The identification and exemplification of multi-word units within a technical corpus of English, including an investigation of nominal groups. Unpublished M.Sc. thesis**,** Aston University.

Tribble, C. (1997). Improvising corpora for ELT: Quick and dirty ways of developing corpora for language teaching. Proceedings of PALC, Retrieved October 2, 2008 from http://www.ctribble. co.uk/text/Palc.htm

Triki, M. (2002). Pragmatics for ESP Purposes. *GEMA Online® Journal of Language Studies. Vol. 2*(1). Retrieved March 11, 2016 from http://ejournal.ukm.my/gema/article/view/218

Waring, R. & Nation, I. S. P. (1997). Vocabulary size, text coverage, and word lists. In Schmitt, N. & McCarthy, M. (Eds.). *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Wilkinson, M. (2014). Using the Keyword Tool to Explore Lexical Differences between British and American English in Specialised Corpora. *CALL-EJ. Vol. 15*(1), 53-70. Retrieved March 11, 2016 from http://www.scopus.com/inward/record.url?eid=2-s2.0-84894852909&partnerID=tZOtx3y1

## ABOUT THE AUTHORS

Noorli Khamis is a Senior Lecturer at Universiti Teknikal Malaysia Melaka (UTeM). She has 20 years of experience in teaching English at different institutions in Malaysia. Her academic achievements include a B.Ed. TESL, M.Ed. TESL and PhD in English Studies. Her research interests are corpus linguistics and ESP.

Imran Ho Abdullah is a Professor of Cognitive and Corpus Linguistics at the School of Language Studies and Linguistics, UKM. His research interests are corpus linguistics and cross-cultural semantics/cognitive semantics and the natural extension of these interests to translation and the use of corpus methodologies in Translation Studies.

eISSN: 2550-2131
ISSN: 1675-8021