VISTAS

Journal of Humanities & Social Sciences

Volume 7 / 8 *2011/ 2012*

Identifying Threshold Vocabulary for Academic Study

Shivanee Ranjula Ilangakoon
University of Colombo

ABSTRACT

Many Asian countries are moving towards English medium instruction in tertiary education (Fan, 2001) because English is gaining currency as a global language. However, studies have revealed student face problems in coping with the demands of their studies due to lack of language proficiency that stems from poor vocabulary knowledge. Within this context it is heartening that vocabulary research posits that 80% of academic texts are covered by the 2000 high frequency words (Coxhead and Nation 2001), 10% by the 570 academic words given in the Academic Word List by Coxhead (2000) and 10% by low frequency or technical words. Nonetheless, there is a question as to the possibility of one core academic word list serving students from different disciplines. Another significant issue that needs to be addressed in terms of vocabulary knowledge is that meaning in texts is derived through word partnerships or collocations ((Lewis 1997b).

The current study was carried out to investigate the feasibility of the idea that one core academic word list can serve students following different disciplines. It also aimed to find keywords or subject specific words and collocations in the corpus.

The Arts subject stream was selected as a case study. An academic corpus, comprising texts from Sociology, Economics, History and International Relations, Political Science and Demography was compiled.

The findings revealed the occurrence of 548 out of the 570 words in the Academic Word List, in the specialized academic corpus that was compiled. The search for keywords that occur in the five sub corpora revealed around 350 keywords in each of the sub-corpora that characterize the vocabulary of each corpus. Many significant collocations were also extracted. These findings emphasized the relevance and validity of the words in the Academic Word List and reveal that awareness of discipline-specific keywords and collocations are also significant in forming the threshold vocabulary requirement for academic study[1].

Key Words: Academic Corpus, Academic Word Lists, Vocabulary, University Education

[1] This paper forms part of the thesis submitted in partial fulfillment of the MA in Teaching English as a Second Language at the Post Graduate Institute of English, Open University of Sri Lanka, Nawala, 2011.

1. INTRODUCTION

Many Asian countries including Sri Lanka are moving towards English medium instruction in secondary and higher education even though there is "a growing concern whether students are proficient enough in English" (Fan, 2001, p. 70) to receive such instruction. As vocabulary development is essential in helping students study in a second language situation, the current study aims to identify vocabulary that can be imparted to students pursuing higher education in order to improve their English proficiency within a relatively limited time period.

Imparting vocabulary knowledge in language courses has regained importance not only because it relates to "all four language skills" (Jordan, 1997. p. 149) and carries the "basic information load" (Read, 2004, p. 146), but also because studies have established that vocabulary proficiency is an indicator of language proficiency (Fan, 2001; Coniam, 1999).

This raises the crucial question of what vocabulary should be taught to undergraduates. While vocabulary research has shown that there are high frequency words that occur in language, the current thinking is that students should be familiar with the most frequent 2000 words in the language that have been posited in the General Service List (GSL)2 by Michael West in 1953 (Coxhead and Nation 2001, p. 252). However, students following English for Academic Purposes (EAP) courses need to go beyond this level and be familiar with academic words. Such words occur in academic texts and are inter disciplinary in nature and have been given in the University Word List (UWL)[3] developedby Xue and Nation in 1984 (Nation 1990 cited in Coxhead and Nation 2001, p. 253) and the Academic Word List (AWL)[4] by developed by Coxhead in 1998 (p. 254). While these lists consist of individual words, Lewis (1997b, p. 254) suggests that native speakers possess a "repertoire of multiwords" at their command. Multiwords are treated as independent units and they are socially sanctioned groups of words. These units could be collocations such as "to raise capital" or institutionalized utterances that occur mostly in speech such as " I'll get it. There's a call for you. If I were you …" (Lewis 1997b, p.257). Coady (1997) looks at research on collocations and affirms that "multiword phrases are not learned well through ordinary language experience and suggests that there is a need for them to be learned explicitly" (p. 282).

The current study was undertaken with a view to establishing threshold vocabulary that students in the Faculty of Arts, University of Colombo, Sri Lanka would find beneficial in

[2]GSL can be downloaded from http://jbauman.com/gsl.html

[3]Uwl can be downloaded from http://jbauman.com/UWL.html

[4] AWL can be downloaded from http://www.uefap.com/vocab/select/awl.htm

their academic careers. This was especially important in the context that the Faculty was contemplating moving over to English medium instruction from mother tongue instruction. Furthermore, as the majority of the students in the Faculty of Arts are from the remote parts of the country, where exposure to English is minimal, their competency level is unsatisfactory. Studies undertaken by Perera (2006), Fonseka (2008) and Kumara (2009) to establish the English language proficiency of undergraduates at the Universities of Kelaniya, Ruhuna and Rajarata respectively have revealed that the competency of a majority of the students is below average.

The university benchmarks designed by the English Language Teaching Units (ELTU) of Sri Lanka attempt to create a set of common criteria to evaluate student performance levels. This was in response to the observation that there is a "wide discrepancy in student standards of English at entry" (Medewattegedara, 2011) among the universities as well as between their respective faculties. The benchmarks, designed as "can do" statements indicate what students at each band of the benchmarks can do. However, they give no specific indication as to the vocabulary knowledge the students should possess. It should be noted, however, that benchmark two refers to "a few frequency words" although no further clarification is made to what it should include.

A study by Kumara (2009) attempted to "recognize the most frequent lexis and structures, which can be incorporated into the syllabus and teaching of the target English language situation" (p. ii). However, the study looks exclusively at spoken texts and these texts are selected from the Faculty of Applied Sciences.

Therefore, it was felt necessary to investigate recurrent vocabulary students would encounter. These vocabulary items would strengthen their language ability as they will soon be called upon to study in the English medium. In order to achieve the above objectives, this study selected five disciplines in the Faculty of Arts for investigation. While written texts in these five disciplines were collected no spoken texts or lectures were collected. The texts were collected in electronic form to compile the corpus which was then analyzed for academic words and subject specific words or keywords and multi word phrases. The analysis of the corpus was carried out using the corpus software tool Compleat Lexical Tutor. The study aimed to make a record of the most relevant academic words for undergraduate students.

2. LITERATURE REVIEW

2.1 Relevance of vocabulary in language teaching

The significance given to vocabulary in English Language Teaching (ELT) has undergone many changes over the years. While early approaches dealt with explicit teaching and rote learning they later gave way to incidental learning, emphasizing "inference and prediction drawing on general word knowledge" (Goodman, 1967 in Cobb and Horst, 2001, p. 316). The pendulum now seems to have swung back midway to incorporate teaching of vocabulary through implicit and explicit exposure (Sokmen 1997). Cobb and Horst's (2001) viewpoint on vocabulary, after conducting research with Omani university students, is that a lack of vocabulary becomes the chief cause for students' weak language aptitude in coping with their studies. Fan's (2001) premise is that there is a positive relationship between language proficiency and knowledge of vocabulary.

Nation and Waring's (1997) search for a learner threshold of vocabulary size has revealed the following information. Out of 54,000 word families in English (Webster third edition) an adult native speaker may possess a vocabulary of 20,000 while 3-5,000 words suffice for basic comprehension, and a vocabulary of 2-3000 words for productive use in speaking and writing. The encouraging information that Nation and Waring posit is that a small number of high frequency words form "a very large proportion of running words in unsimplified texts" enabling a learner with a vocabulary of 3000 words to function effectively in the language. High frequency words are those that occur most frequently in a language. Cobb and Horst (2001) endorse the concept that a learner could have a 2000 word range and command comprehension of 80 per cent of a text (2001). Both studies consider the learning of high frequency words as a hopeful message for learners. Nation and Waring (1997) point out that once the first 2000 high frequency words are learnt the next level of words will depend on the needs of the learner. If the need is a social one the vocabulary required will be low frequency words while the need to pursue academic study at university level will require the next level of vocabulary "variously called semi-technical, sub-technical or academic, and consists of words which occur across a number of disciplines" (Jordan, 1997, p. 152). Thus academic vocabulary is beyond the 2000 high frequency words and is interdisciplinary in nature and occurs in academic texts of different disciplines. High frequency words and academic words have been identified in different word lists as discussed below.

2.2 Word lists

Nation and Waring, (1997) make a detailed study of word lists that were used at the time of their research. They discuss four word lists giving special emphasis to the General Service List (GSL) developed by Michael West in 1953 which comprises the most frequent 2000 words in the English language, and provides 82 per cent coverage of written texts. They advocate the University Word List (UWL), containing over 800 word families for students at tertiary level who wish to go beyond the basic level. They consider the frequency lists valuable in setting "learning goals" in "curriculum design", and assist the teacher in vocabulary and text selection to focus on in classroom teaching (p. 17). Subsequently, Coxhead (2000) made a significant contribution to lists with her Academic Word List (AWL) that consists of 570 word families from a range of disciplines. Words that were selected to formulate the AWL were outside the most frequent 2000 words.

2.3 Creation of word lists through corpus compilation

It is important to consider what principles guide the formation of these word lists and how reliable they are. Linguists have formed vocabulary lists based on data obtained from collections of written and spoken data or through the compilation of a corpus.

Corpus linguistics, a comparatively new field in the study of language, investigates "actual patterns of language" in both spoken and written texts (Reppen and Simpson, 2002, p. 92). For this purpose linguists make "a large principled collection of natural texts" (Reppen and Simpson, 2002, p. 93), called "a corpus" (plural form-corpora).

Corpus compilation and analysis have progressed tremendously with the use of computers that are able to process text data in electronic form. As texts can be typed, copied, scanned or simply downloaded from the world-wide-web, text collection can be performed rapidly and with ease (Reppen and Simpson, 2002). An even greater facility is provided by word processors through software packages that can be used to effortlessly access and analyze the stored data in numerous ways to observe endless linguistic phenomena. High frequency word lists, keyword lists, multiword phrases and collocations are some of the features that can be observed through such analysis. The features relevant to this study are discussed below.

Hunston defines keywords as "words which are significantly more frequent in one corpus than another" (2002, p. 68). In order to search for keywords a specialized corpus has to be compared with a general, bigger corpus using software. Hunston (2000) further states that keywords can be 'lexical items' and they will indicate topic specific vocabulary but they can also include grammatical items that will reveal more specific information about the language preferences of a particular discipline. Kumara (2009) in his study searched for key words in the full corpus as well as sub corpora in order to "recognize the specialized lexis in each case" (p. 40).

Coxhead and Nation (2001) contend that academic vocabulary is of immense value as it is "common to a wide range of academic texts" and accounts for around 10% of an academic text (p. 254). They go on to point out that academic vocabulary is less known than technical words but is the "kind of specialized vocabulary that an English teacher can usefully help learners with" (p. 256).

In order to posit such a word list Coxhead gathers data from a well designed collection of texts of academic English (Coxhead, 2000). This list is created out of an electronic collection of 3.5 million words which "involved 28 subject areas organized into 7 general areas within each of the four disciplines; arts, commerce, law, and science" (p. 216). The Academic Word List (AWL) created out of this collection presents 570 words.

Rejection of one core academic word list.

However, Hyland and Tse, (2007) question the validity of Coxhead's AWL by juxtaposing it with the results of their own corpus. Their research suggests that "individual lexical items on the list often occur and behave in different ways across disciplines in terms of range, frequency, collocation and meaning" (p. 235). They argue that the behaviour of words differ in subject streams such as science, law, and humanities and that homography or single word forms with different meanings can "misrepresent the composition of word families" (p. 243). While they question the existence of one word list for students of all disciplines due to the above reasons they suggest that teachers help students develop their own subject specific word lists. However Cobb and Horst (2001) are of the opinion that "one frequency based threshold" is more useful than "a separate threshold for every text" (Cobb and Horst 2001, p. 318).

2.4 Collocation

Yet another aspect that can be searched with the use of corpus analysis software is collocation. Hunston (2002) defines collocation as the "statistical tendency of words to co-occur" (p. 12) and points out that corpus-based concordance lines can be of immense value to statistically bring out collocations in particular text domains (Hunston, 2002, p. 68). A programme designed to present collocations will show the node word in the center of the screen with the words it collocates with to the right and left of the selected word within a particular span (Hunston, 2002, p. 68). While it is possible for some of these words to appear "by chance", others are considered "meaningful". In her search for the word 'gaze' Hunston posits that the words "penetrating gaze/ my/ her/ his gaze" to be 'meaningful" while "wait" and "life" to be two terms that occur by chance (2002 p. 69).

Nattinger and DeCarrico, (1992) state that if "the node word occurs with a span of particular words at a frequency greater than chance would predict, then the result is collocation" (p. 20). Furthermore, they state that large corpora are now developed to observe how language functions in practice. 'Whereas syntax deals with general classes of words and their combinations, collocations describe specific lexical items and the frequency with which these items occur with other lexical items (p. 20).

They make a distinction between syntactic strings, collocations and lexical phrases. They define collocations as "strings of specific lexical items, such as rancid butter and curry favor, which co-occur with a mutual expectancy greater than chance. These strings have not been assigned particular "pragmatic functions" (p. 36). While syntactic strings are produced through "syntactic competence" lexical phrases are collocations that "have been assigned pragmatic functions" such as "how do you do" (p. 36).

The current study aims to ascertain the validity of the AWL through the compilation of an academic corpus. This corpus will be searched for academic words, keywords and collocations that will be of relevance to vocabulary development of undergraduates.

3. METHODOLOGY

3.1 Corpus design and compilation

Many "principled decisions" need to be taken in order to design and compile a corpus, due to the fact that the design "impacts all of the analysis that can be carried out with the corpus and has serious implications for the reliability of the results" (Reppen and Simpson 2002 p. 93). While key issues in corpus compilation deal with representativeness, balance, size and content of the selected texts, the research questions play a pivotal role in guiding the decisions regarding the above issues. Thus while the following research questions framed the design of the corpus titled Arts Academic Corpus (AAC), much care was taken to balance the size of the sub corpora and select representative samples of texts to investigate the selected subset of the language.

Research Questions

1. How does an academic word list created out of the Arts Academic Corpus compare with Coxhead's Academic Word List (1998)?
2. What are the keyword lists created out of the Sociology, Demography, History and International-Relations and Economics sub-corpora?
3. What are the subject specific collocations that can be extracted from the five sub corpora?

The research is carried out on the principles guiding corpus linguistic studies approach of corpus compilation and analysis. Its more generic research is a Case Study of the Faculty of Arts of the University of Colombo.

As the corpus was meant to identify the academic words that the undergraduates of the Faculty of Arts in the University of Colombo would encounter, the first step was to select written academic texts from the subjects the students follow. Although the Faculty of Arts offers ten subject streams, only five were selected for the purpose of this research. Written texts from Sociology, Demography, Political Science, History and International Relations and Economics were selected to compile the corpus.

3.2 Data analysis tools and their features used in the study

As corpus based studies are gaining popularity, the need for software programmes that are user friendly and efficient are also being developed rapidly. There are many software programmes currently available in different modes. For the purpose of the current study, the software program The Compleat Lexical Tutor (CLT) was utilized as it best suited the research questions. The CLT created by Tom Cobb is an online software programme offering a number of tools and sites and can be accessed on http://www.lextutor.ca/ .Three different types of word lists were obtained from the AAC using the above programme.

The feature 'VocabProfiler' in the CLT breaks texts down into words according to word frequencies. All words in an uploaded text or corpus will be broken down into the 'first thousand frequent words', 'second thousand frequent words', 'academic words' and the remainder as 'offlist' words. Using the CLT, the study aimed to create an academic word list in order to identify the number of academic words in the corpus and to identify what they are. Although it was possible to find the academic words in the sub-corpora, the researcher encountered an obstacle in obtaining the results of the full corpus. As the software was unable to analyze the full corpus due to its size it became necessary to recompile the corpus and limit its word count to 100,000. The original number of articles was retained but the words per article were reduced to approximately 2000. The recompiled corpus was named version B of the Arts Academic Corpus (AAC), while the original corpus was named version A of AAC.

In order to identify the key words of the AAC and its sub-corpora the 'Keyword Exractor v.1' of the CLT was utilized.

Collocation was investigated through concordance lines which display multiple occurrences of a word that is being searched. The search for collocations was carried out to identify discipline specific collocations that occur in the five sub corpora by using the tool 'N-Gram extractor' of the CLT.

4. FINDINGS

4.1 Arts Academic Corpus

The coverage of the academic words in the AAC is 11.11% while the academic words in its sub-corpora range from 9.50% - 12.44% indicating that AAC is truly academic in nature as an academic corpus should contain 8-10% academic vocabulary in its corpus (Nation, 2001 cited in Hyland and Tse 2007, p. 236). The percentages of academic vocabulary validate the corpus as worthy of creating an academic word list that could be utilized for pedagogic purposes and for comparing it with the Coxhead word list.

In Table 01 the statistics of Version B of the corpus in relation to the 'first thousand words', 'second thousand words', the 'academic words' and the remaining words as 'offlist' words are given.

Table 1 Statistics of the word categories in version B of AAC (100,000 words)

Families	Types	Tokens		Percentage
K1 Words (1-1000):	898	2415	73052	73.01%
Function:			(41594)	(41.57%)
Content:			(31458)	(31.44%)
K2 Words (1001-2000):	535	955	4610	4.61%
1k+2k				(77.62%)
AWL Words (academic):	548	1551	11111	11.11%
Off-List Words:		4067	11280	11.27%
	1981	8987	100053	100%

4.2 The academic words from the AAC

In order to answer Research Question 1, Version B of the corpus was utilized to extract the academic words list from the AAC using CLT. This progamme provides two lists: one which comprises 548 types or the head word list of the corpus and the second list provides the families of the 548 academic words with their number of occurrences.

When comparing the academic words extracted from the current corpus with Coxhead's AWL it was seen that while Coxhead's Academic Word List consists of 570 words created out of a 3.5 million running words the AAC consists of 548 academic words created out of a 100,000 word corpus. Although the two corpora are very different in areas of the sub corpora, date of publication and size an overwhelming majority of words have occurred in both lists. With this observation in mind, it is worthwhile to examine the criticisms against the AWL which initially prompted the researcher to carry out this project.

Criticisms against the Coxhead Academic Word List

While Hyland and Tse (2007) question the acceptability of "a single core vocabulary for academic study" (p. 235) they express concern over one list that can serve students of different disciplines. As the large majority of words in the two word lists overlap in spite of the huge difference in the size of the corpus and the composition of the corpus, the current list seems to suggest that the AWL is an acceptable academic word list that can be utilized to teach students in different academic disciplines. Thus all words in the list were felt to be of importance for academic study and are justified in being included in the list.

4.3 Keywords

Considering the fact that different disciplines will have lexical preferences the CLT was utilized to search for 'keywords' or words that are specific to a particular discipline which formed the second research question. The keywords occur with an unusually higher frequency in the specialized corpus when compared with a much larger general reference corpus (Bowker and Pearson, 2002). The reference corpus provided by Compleat Lexical Tutor and utilized in this study is the Brown Corpus which has 1 million words drawn from 500 texts.

Through the use of the CLT the following number of keywords was extracted from the subcorpora: 345 keywords from demography, 424 keywords from Economics, 376 keywords from History and International Relations, 355 keywords from Political Science and 359 from Sociology. After a manual sorting of these words the researcher would like to present 30 keywords each from the five disciplines and are given in Table 02.

Table 2: Selected keywords in the sub-corpora

Demography	Economics	History and INR	Political Science	Sociology
1. fertility	currencies	global	negotiation	affluent
2. migration	households	bilateral	federalism	motivation
3. abortion	bilateral	legitimacy	arbitration	citizenship
4. aging	commodity	axioms	conflicts	interactions
5. mortality	aggregate	idealism	propositions	behaviour

Demography	Economics	History and INR	Political Science	Sociology
6. expectancy	monetary	dictate	stalemate	illegitimacy
7. households	consumption	intervene	disarm	syndrome
8. births	economies	outbreak	unilateral	ethnic
9. determinants	substitution	consensus	conducive	communities
10. migrants	equilibrium	ideologies	maritime	parental
11. gender	reductions	beliefs	resolving	challenges
12.demographic	strata	nationalist	disagreements	segregation
13. pyramid	stabilize	Reich	neutrality	urban
14.maternal	protocol	imperialism	advocated	caste
15. populations	instability	elite	legislatures	occupations
16. decline	inflation	emergence	empires	vulnerability
17. prevalence	asymmetric	territorial	autonomy	accessibility
18. regimes	influencing	realist	grievances	alleviation
19. projected	investment	capitulation	disputed	instability
20. dynamics	output	regimes	ecological	byproduct
21. longevity	determinants	conflicts	settlement	migrants
22. fertile	incentives	industrialization	mainstream	stabilize
23.transition	global	collapse	factions	sanitary
24. famine	curve	traditionalism	adversaries	impacted
25. epidemics	regimes	monarch	liberation	disability

Demography	Economics	History and INR	Political Science	Sociology
26. declining	subsidies	asymmetric	contested	prosper
27. stabilize	fundamentals	sovereignty	mistrust	gangs
28. aggregate	export	empire	talks	intervene
29.elderly	labour	treaty	correlated	spouse
30. deaths	impacts	pluralistic	endorsement	fostering

4.4 Collocations from the sub corpora.

The third Research Question sought to find subject specific collocations. In each sub-corpora it was found that there were collocations that are discipline-specific as well as collocations that a student of any discipline could use. Discipline specific collocations are shown in Table 03. These collocations were selected on the basis of their high frequency and the usefulness for pedagogical purposes identified by the researcher. While the lists are by no means exhaustive it is expected to raise awareness in the students of the existence of such units from which meaning needs to be derived. Lewis (1997a) is of the opinion that by raising awareness of collocations, students "communicative power" can be improved "even with the limited language resources at their disposal" (p. 33). In addition, Lewis (1997b) stresses their value by stating that "[n]ative speakers, in addition to words and grammar, have at their disposal a repertoire of multi- word items that are, for certain purposes, treated as independent units" (p. 255).

Table 3: Collocations created out of the sub-corpora

Demography

Subject specific Collocations	
1.	life expectancy at birth
2.	the age structure of
3.	growth rate of the population
4.	the small birth cohorts of
5.	the standard of living
6.	population growth
7.	maternal and child health
8.	labour migration
9.	segment of the population
10.	census and statistics
11.	transition from the Malthusian regime
12.	income per capita
13.	induced abortion
14.	less developed countries
15.	side effects

Economics

Subject specific Collocations	
1.	goods and services
2.	aggregate demand
3.	regimes for major currencies
4.	currency exchange rate
5.	closed economy
6.	the housing bubble
7.	exchange rate
8.	the demand for
9.	in a plural society
10.	per capita income
11.	equilibrium in the goods
12.	low and middle income countries
13.	FDI flows to developing countries
14.	the rest of the world
15.	the production possibility

History and International Relations

	Subject specific Collocations
1.	the cold war
2.	foreign policy
3.	regional cooperation
4.	outbreak of the first world war
5.	territorial sovereignty
6.	national self determination
7.	balance of power
8.	South Asian countries
9.	ethical principles
10	central arguments in
11.	public sphere
12.	regional economic integration
13.	the British empire
14.	the diversity and complexity of
15.	the right time

Political Science

	Subject specific Collocations
1.	the parties to the conflict
2.	the role of the mediator
3.	come to the negotiation table
4.	failed negotiation
5.	was elected the
6.	to preserve the territorial integrity
7.	unitary structure of the state
8.	resolve the conflict
9.	the parties involved
10.	negotiated settlement
11.	under the leadership of
12.	the main advocates of the
13	get what they want
14.	demand for
15	to change the course of

It can be noted that phrases such as drop out, negotiation table, per capita income, life expectancy, standard of living can be considered as fixed phrases where it is unlikely that a word in the phrase would differ depending on context of use. Lewis (1997b) refers to them as fully fixed collocations while acknowledging that there are yet others that can be relatively fixed. In the phrase 'the process of globalization', the word globalization could be replaced by another. Similarly, in the phrase the 'role of the mediator', mediator could be replaced. While looking at the lists given in order of frequency, the selection given here is according to its usefulness in teaching collocations. The phrase 'the nature of', 'demand for', 'was elected the', are part of a full phrase that could be completed with different words. The given collocations can be taught to raise awareness in students that meaning needs to be derived from a phrase rather than from an individual word. Being aware of this phenomenon can aid receptive skills and enrich language production.

5. CONCLUSIONS

5.1 Findings from word lists

In the current study, the researcher looked at the Academic Word List, the key word lists and collocations that were generated from the AAC and its sub-corpora to verify their significance in terms of pedagogical relevance to the undergraduates of the Faculty of Arts. The following observations were made regarding the three sets of lists.

5.1.2. Academic word list

As the AWL was created out of the AAC version B, which is a corpus of 100,000 words, containing 548 of the words in the 570 Academic Word list, it can be concluded that the word list by Coxhead is of great significance for classroom teaching of academic vocabulary. The fact that the 548 words were extracted out of a totally different and much smaller corpus indicates the likelihood of these words occurring in different academic disciplines. Although some of the words in the list are less frequent than others, students should be exposed even to these less frequent items, because as pointed out by Coxhead (2000) the less frequent items have the tendency of being overlooked by students. Furthermore, due to their infrequency of occurrence the possibility of incidental learning taking place is also remote. As academic words are considered to be of a more advanced nature than the first 2000 words (Nation and Waring 1997), these words would certainly be appropriate for undergraduates at band 6 and 7 of the University benchmarks as the mandate of these two bands is to create a learning situation to impart academic language. Furthermore, as the university benchmarks do not specify the vocabulary to be taught at this level, the researcher would like to recommend that the AWL vocabulary be taught to students at benchmarks six and seven. Overall, the benchmarks do not specify or identify any particular vocabulary to be taught to undergraduates although benchmark 2 refers to "a few high frequency words". As such the first two thousand most frequent words should be taught to students at benchmarks 1 to 5. Coady (1997), making a synthesis of vocabulary acquisition research, "emphasizes that these words should be learnt to the point of automaticity" (p. 279). He also supports Nation's point of view that the most frequent 2000 words should be learnt as soon as possible through efficient methods that include both "direct teaching and learning and the use of graded readers" (Coady, 1997, p. 279).

The AWL and the word families of the AWL are two lists that provide naturally occurring data to teach word classes to students at different levels. Students could be taught affixation along with word classes which would result in students accumulating a much larger vocabulary than the 570 headwords in the AWL. As recommended by Coxhead, if students are aware of word building strategies they will require little effort to learn the derived forms of a root word (2000). Thus the 570 words in the AWL can be used to teach these words as well as their derived forms.

5.1.3 Keyword lists

The keyword lists demonstrate the subject specific vocabulary in the corpus. When scrutinizing the key word lists, it becomes evident that all words in the list need not be selected for teaching in an EAP course. Some of the words were proper nouns that were of high frequency in the corpus while there were also words such as shirts and tailor. These words could be removed from the list and the remaining keywords could be used as vocabulary items that could be incorporated into language lessons. These words could be used in setting vocabulary goals or targets within a language course. These lexical items could also be included in the glossaries that departments prepare for their students. The proper nouns that occur in the lists are either key figures in the different disciplines or researchers or authors of texts. Key figures such as Malthus, Marx, Stalin, Hitler, Lenin (names which appeared in the corpus) could be used as reading material for students from band one to four to teach students how to describe people. The places that figured such as Vietnam, Asia, Afganistan, Pakistan and Iran in the History and International Relations sub corpora could also be used in teaching about places at the same band levels.

The words identified can also be used in selecting EAP reading material to be used in the language classroom. This will make it possible to focus students' attention on the identified words in context. Contextual exposure of vocabulary will aid retention better (Read 2004 and Sokmen, 1997) while additional post-reading activities on the selected items will provide repeated exposure that is vital for retention of newly acquired vocabulary items (Hyland and Tse, 2007).

5.1.4 Collocations

Many lexical patterns that are subject specific and general in nature were discovered and a selected number are recorded in this study. The current study identifies collocations that are subject specific and could be taught for their meanings as well as to raise awareness of their occurrence in academic texts. Making students aware that meanings in texts are derived from multiword units which may be different from the individual meanings of the words in the phrase is an important aspect of vocabulary.

5.2 Pedagogical implications

Having identified the vocabulary students are bound to encounter, the next question is how best it needs to be imparted to the students. While mere lists alone cannot be given to students with the expectation that they will be memorized, a multi pronged approach should be adopted in imparting the identified vocabulary in order to make learning possible. The lexical approach recommends many useful vocabulary activities and strategies (Lewis 1997a) that can be used to impart lexical knowledge. At the same time there is no denial that contextual input (Read 2004; Sokmen, 1997) is the best method for vocabulary development. Word lists that had once being discarded have regained their significance (Nation and Waring 1997; Sokmen, 1997). Keeping these studies in mind, this researcher would like to recommend a mixed strategy for utilizing the vocabulary lists in the classroom. A selected section of the academic word list, the keyword lists and collocations should be taught through vocabulary recognition activities. Another set or the same set should be given to students to engage in self learning through dictionary or web searches and making a record of such findings. The search can include meanings of words, word classes, collocations and words in sentences. These findings could go into creating students' vocabulary note books. The third strategy would be to ensure that students encounter the selected items in context in the classroom. This has implications for the selection of reading material rich in the identified vocabulary. Authentic texts could be used for this purpose with modifications that involve multiple occurrences of the target vocabulary. Finally the full lists should be given to the students and they should be made aware of the vocabulary from which they can benefit in their academic pursuits. Learners should also be told to be conscious of these words and pay attention to how they are used by writers and use the above strategies in discovering meanings, patterns and usage of items in the list that interest them.

The above recommendations are made with two major objectives in mind. The first is to make students aware of the vocabulary items that are most important due to their frequency of occurrence in academic texts. Being armed with knowledge of such lexical items will facilitate academic reading comprehension. Qian (1998) posits that vocabulary size and depth are closely related to reading comprehension according to the results of his study of general academic text comprehension of 74 adult Chinese and Korean speakers. The second objective is to ensure repeated exposure to the selected items as retention of vocabulary requires repeated exposure to them Cobb (1997). Lewis (1997a) suggests as many as seven times of encountering the item, which is not necessarily through explicit teaching, for acquisition to take place.

FUTURE RESEARCH

The current study has attempted to record the academic words, keywords and collocations found in a written corpus, and to point out some possible strategies that could be used in teaching them. However, much more research will have to be carried out on best practices and strategies to be employed in imparting the identified vocabulary to the students. The strategies students adopt in vocabulary learning also form an important area for future research. Furthermore, research should also be carried out on investigating lexical patterns in spoken academic texts.

REFERENCES

Bowker, L. & Pearson J. (2002). *Working with specialized language: A practical guide to using* corpora. London: Routledge.

Coady, J. (1997). L2 vocabulary acquisition: A synthesis of research. In J. Coady, and T. Huckin, (Eds.), *Second language vocabulary acquisition* (pp.273-290). Cambridge: Cambridge University Press.

Cobb, T. & Horst, M. (2001). Reading academic English: Carrying learners across the lexical threshold. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp.315-329). Cambridge: Cambridge University Press.

Cobb, T. (1997). *From concord to lexicon*: *development and test of a corpus-based lexical tutor*. Doctoral dissertation, Concordia University, Montreal, 1997. Retrieved from http://www.lextutor.ca/cv/webthesis/Thesis0.html

Coniam, D. (1999). Second language proficiency and word frequency in English. *Asian Journal of English Language Teaching*, 9, (59-74)

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

Coxhead, A. & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew and M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp.252-267). Cambridge: Cambridge University Press.

Fan, M. Y. (2001). An investigation into vocabulary needs of university students in Hong Kong. *Asian Journal of English Language Teaching*, 11, 69-85. The Chinese University Press.

Fonseka, G. (2008). The teacher's role in achieving student empowerment through English. In D. Fernando, & D. Mendis, (Eds.), *English for equality, emploment and empowerment* (pp. 27-37). Colombo: Ceylon Printers Ltd.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K. & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly*, 41, (pp. 235-253)

Jordan, R.R. (1997). *English for academic purposes-a guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kumara, M.D.S.S. (2009). *Compilation and linguistic analysis of a dedicated corpus for the Applied Sciences: Special focus on the spoken academic discourse of the applied sciences study programme of the Rajarata University of Sri Lanka*. Unpublished MA dissertation, University of Kelaniya, 2009.

Lewis, M. (1997a). *Implementing the lexical approach: Putting theory into practice*. Cengage Learning.

Lewis, M. (1997b). Pedagogical implications of the lexical approach in second language vocabulary acquisition. In J. Coady, & T. Huckin, (Eds.), *Second language vocabulary acquisition* (255-271) Cambridge: Cambridge University Press.

Medawattegedara, V.V. (2011). *English and English language testing HETC project*. Paper presented at University Development Grants (UDGs) Training Program for Proposal Writers, Ministry of Education, Colombo, February.

Nation, P. and Waring, R. (1997). *Vocabulary size, text coverage and word lists*. In N. Schmitt, & M. McCarthy, (Eds.), Vocabulary: Description, acquisition and pedagogy (pp.237- 257) Cambridge: Cambridge University Press.

Nattinger, J. R. and DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Perera, K. (2006). Laying the foundations: Language planning in the ELTU. In H. Ratwatte, & S. Herath, (Eds.), *English in the Multilingual Environment* (pp. 47-57). Nawala: Open University of Sri Lanka.

Qian,D. D. (1998). *Depth of vocabulary knowledge: Assessing its role in adults' reading comprehension in English as a second language*. Ph. D Dissertation, University of Toronto. Retrieved from http://www.collectionscanada.gc.ca/obj/s4/f2/dsk2/ftp02/NQ33914.pdf

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, 146- 161. Cambridge: Cambridge University Press.

Reppen, R. and Simpson, R. (2002). Corpus Linguistics. In N. Schmitt, (Ed.*), An introduction to applied lingu*istics (92-111). London: Arnold.

Sokmen, A. J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* Cambridge: Cambridge University Press.