ISSN: 0128-7702
Pertanika J. Soc. Sci. & Hum. **18 (2)**: 241 - 258 (2010)

Analysing Collocational Patterns of Semi-Technical Words in Science Textbooks

Sujatha Menon[1]* and Jayakaran Mukundan[2]
[1]Academy of Language Studies, Universiti Teknologi MARA,
Kampus Bandaraya Melaka, 110, Off Jalan Hang Tuah,
75300 Melaka, Malaysia
[2]Department of Language Education and Humanities,
Faculty of Educational Studies, Universiti Putra Malaysia,
43400 UPM, Serdang, Selangor, Malaysia
[*]E-mail: sjathark@yahoo.com

ABSTRACT

This paper analyses the discourse of science through the study of collocational patterns of semi-technical words in science textbooks used by upper secondary students in Malaysia. Semi-technical vocabulary is considered to be one of the most problematic lexical areas for second language learners learning science in English, as these words usually take on extended meanings in technical contexts. The study explores the collocational and colligational patterns of semi-technical words in a corpus of 12 science textbooks. The analysis found some common lexical and grammatical patterns which seemed to share similar aspects of meaning and also showed that collocations of semi-technical words often form compounds with extended and more genre specific meanings. Thus, it is important to focus on some sense and grammar patterns as students need to understand why even though many of these collocations share similar syntactic characteristics, the flexibility of some combinations are arbitrarily blocked by usage.

Keywords: Semi-technical, corpus, EST, collocation, prescribed textbooks, ESP

INTRODUCTION

Insights from corpus research have revolutionized the way language is viewed, especially words and their relationship with each other in context (Schmitt, 2000). The aim of corpus linguistics is to analyse and describe the language use as realised in the selected texts (Tognini-Bonelli, 2001).

Corpus linguistics does not begin by accepting certain rules as given, in fact, it defines its own sets of rules before being applied and provides new rules and parameters for linguistic description (Tognini-Bonelli, 2001). Corpus research starts from the assumption that meaning, in its different forms, is realized foremost at the linguistic level (Firth, 1957 in Tognini-Bonelli, 2001, p.157). Tognini-Bonelli (2001) adds that corpus research allows us to investigate the co-selection of words or the way in which words interact with each other as these patterns of interactions between words can create new and complex units of meaning.

There is reasonable consensus that a corpus will not just provide insights into the contents but also that the results of the analyses will be claimed to be typical of the language from which the corpus was selected. Corpus research allows researchers and learners to gain insights into the language, particularly the interconnection of lexical and grammatical patterns, collocations,

Received: 24 June 2009
Accepted: 18 June 2010
[*]Corresponding Author

colligations, the frequency of words and the use and functional behaviour of these words (Tognini-Bonelli, 2001; Schmitt, 2000; Nelson, 2001; Sinclair, 1991).

The basis of using corpora in language teaching and learning is the rising awareness that learners need to understand that there is variability in language use (Stern, 1992). Rather than relying on a strict code of rules in learning language which quite often restricts input, learners are given ample opportunity to discover language and systematize it for themselves for better understanding and acquisition of language (Tomlinson, 1998; Willis, 1998). Tomlinson (1998) adds that this awareness can make learners more attentive to salient features of their input and can facilitate language acquisition.

Hunston and Francis (2000) insist that through corpus analysis, learners would be encouraged to think of grammar differently in that a corpus can provide the learner and the teacher information about 'what is or is not said in a given language or variety and what certain grammatical choices mean' (2000, p.260). This suggests that grammar cannot explain but can generalize and that existing rules on grammar can be mostly used as an abstract guideline. It would point out frequently occurring features of language and also 'deviant' patterns or language which is not typical but are strongly associated to particular registers (Gavioli, 1997). Awareness of what is typical and untypical will provide the learners with more autonomy to be creative in language (Hunston and Francis, 2000).

THE LANGUAGE OF SCIENCE

The language of Science is far different from the languages that students use in other subjects areas (Laplante, 1997). The language of Science and English for Science and Technology is specialized as scientific enquiry that requires the learners to describe, interpret, and explain various steps in the Science process (Ary, Razavieh and Jacobs, 1985).

Cummins (1981:1979) suggests that to understand any academic subject matter, students have to be more than proficient in their English communication skills. He further explains that students may be proficient in the general conversational English skills, or as defined by Cummins (1981), the basic interpersonal communication skills (BICS), but they may lack the necessary cognitive academic language proficiency (CALP) to learn Science or any other subject matter.

Linguistic theorists (Rosenthal, 1996; Spurlin, 1995; Krashen and Biber, 1988; Cummins, 1979) describe CALP as the type of language proficiency needed to enable students to learn a context reliant subject, which usually relies heavily on oral explanations of abstract and complex ideas, which is often the way in which Science in schools is taught.

Students are likely to benefit from instruction that targets unfamiliar words, expressions and syntax (Laplante, 1997) as ordinary words and phrases, more often assume a different meaning in content subjects (Nation, 2001; Thompson and Rubeinstein, 2000). One of the ways to improve the understanding of Science discourse and texts in the classroom and to learn about the specific sentence structures, lexis and grammar, is to get the students to engage with the data or texts (Tognini-Bonelli, 2001). This would require most immediately, the creation of a specific database or corpus.

EST RESEARCH

There have been many studies on English for Science and Technology (EST) discourse published over the past 50 odd years. Most of them, however, have been centred on the academic discourse in science and technical research articles (Tarone, Dwyer, Gillette and Icke, 1981; Rodman, 1991:1994 cited in Atkinson, 1999, p.197; Myers, 1992 cited in Atkinson, 1999, p.196; Gledhill, 1996; Grabe and Kaplan, 1997 cited in Atkinson, 1999, p.196; Swales, 1998; Marco, 2000; Soler, 2002; Burrough-Boenisch, 2003), with the exception of a few which were carried out on textbook discourse and teaching and learning materials (Barber, 1962; Higgins, 1967; Lackstrom, Selinker and Trimble, 1975;

Kornwipa, Somcheon and Cowan, 2001). Most of these studies have been focused on specific grammatical features or lexical items in written academic discourse, whilst the lexicogrammar patterns and relationships have been analysed by a few such as Barber (1962), Higgins (1967), Haliday and Martin (1993) and Swales (1998), whose studies have shown that the language in science has its own unique lexicogrammatical patterns.

The role and structure of grammatical elements in scientific writing have been the focus of substantial research, such as that carried out by Barber (1962) who examined vocabulary, clause-types and the use of non-finite verbs in texts of different genres, that of Lackstrom, Selinker and Trimble (1975), who studied passive-stative distinctions, modal use and use of noun compounds in various scientific materials, and that of Soler (2002) who examined the use of adjectives in scientific discourse. Other researchers have looked into lexical cohesion strategies (Myers, 1992 cited in Atkinson, 199, p.196), collocations and multi-word units (Master, 1991 cited in Atkinson, 1999, p.196; Gledhill, 1996: 2000; Biber, Conrad and Cortes, 2004) and collocational frameworks (Marco, 2000).

It is no doubt that literature on the study of language used in scientific research articles is substantial and has covered a wide range of areas. However, not much research has been carried out on other types of scientific writing especially, that of scientific English used in textbooks. Even though there have been studies of the verb forms (Barber, 1962; Lackstrom, Selinker and Trimble, 1975) and discourse structure (Higgins, 1967; Trimble *et al*., 1975) in science textbooks, these areas were looked into in only a handful of studies. Meanwhile, little attention has been paid to the use of semi-technical vocabulary in scientific discourse, which according to Trimble (1985) in his analysis of a scientific discourse, is considered to be one of the most problematic lexical areas for students learning science in English.

SEMI-TECHNICAL VOCABULARY

The vocabulary of science and the technical categories of the lexis of science have been discussed and categorized by many linguists. The most notable categorizations of the lexis of science are by Cowan (1974) and Nation (2001). Both Cowan and Nation had similar categories or degrees of 'technicalness' (Nation, 2001). A summary of Cowan's and Nation's categories are described below:

1. Highly technical words – these are words which appear rarely outside its particular field such as 'epithelial' and 'chromosome' in the science and medical fields.
2. Sub-technical words – these are 'context independent' words (Cowan, 1974, p. 391) which occur with high frequency across disciplines but the majority of their uses with a specific meaning are related to this field. The specialized meaning it has in this field is readily understood outside the field, such as the word 'memory' in the computing field (Nation, 2001, p.199).
3. Semi-technical words – these are words which have one or more general English language meanings and which in technical contexts take on extended meanings.
4. Non-technical words – these are words which are common and have little specialization of meaning, for example 'hospital' and 'judge'.

Trimble (1985) believes that non-native learners do not usually have a problem with highly technical vocabulary as it is taught explicitly by content or core subject teachers. However, learners would face difficulty in comprehending semi-technical vocabulary as these words tend to take on extended meanings in technical contexts. Trimble shows the different meanings the word 'fast' (1985, p.130) assumes in two different scientific fields. In the medical field, 'fast' means 'resistant to' while in the mining field it means 'a hard stratum under poorly consolidated ground'. Due to this nature of acquiring extended meanings, Trimble (1985) feels that semi-technical vocabulary has to be given more focus especially for second language learners (non-native) learning science in English.

CORPUS BASED ON TEXTBOOKS

The importance of textbooks, as a component of science instruction, has been advocated by many researchers (Chiapette, Sethna and Fillman, 1991; Gottfried and Kyle, 1992), in spite of the trend to minimize textbook use in some circles (Ansary and Babaii, 2003). The textbook is and has always been an important aspect of teaching in Malaysian schools. It has become indispensable as teachers depend on them for the provision of tasks and tests for students and the students use them as references.

Though there is heavy reservation against using textbook language as corpus data, as the criticism levelled against it is that it is not naturally occurring language, it should be noted that in Malaysia, the students main exposure to the English language is through formal education and prescribed school textbooks (prescribed by the Curriculum Development Centre of the Malaysian Education Ministry). An analysis of the language in these textbooks would lead to a better understanding of the type of language used to teach science. This would then help teachers to teach science in English and also identify the type of language which should be incorporated into the EST textbook (a subject taught in upper secondary science classrooms in Malaysia).

Research on corpora of language teaching textbooks has enabled the examination of the language to which learners are exposed, and when compared to reference corpora or real-language corpora, has resulted in the development of more effective pedagogical materials (Gabrielatos, 2005). The advantage of a pedagogic corpus is that when an item is met in one text, examples from similar previous texts can be used as evidence for the learner to draw conclusions about that language. In other words, it helps learners to recognize patterns or phraseology particular to that discourse (Hunston and Francis, 2000; Sinclair, 1991). Other than benefiting learners and teachers, a pedagogic corpus would be useful in re-designing teaching materials in the future.

This work is part of a larger study on the language used in secondary school science textbooks in Malaysia. As there is no existing corpus of the language used in the teaching and learning of science in schools in Malaysia, the researchers started off by creating a corpus of the language used in all the prescribed science textbooks, which are used in the secondary schools (lower and upper secondary) throughout Malaysia. This paper is based on an analysis of prescribed textbooks of all the major science subjects (General Science, Biology, Chemistry and Physics) taught in upper secondary, form 4 and form 5 (ages 16-17 years) classrooms.

OVERVIEW OF THE PRESENT STUDY

This paper explores corpus evidence on the collocation and colligation patterns of semi-technical vocabulary in the upper secondary Science corpus. As semi-technical words are considered to be challenging for students (Trimble, 1985; Nation, 2001) and is a focus area in the Malaysian EST syllabus, an analysis of these semi-technical words would not only help establish knowledge of the type of words which frequently occur and are associated with semi-technical words, but also help to identify some of the phraseology specific to scientific English used in textbooks.

Collocation in this work is taken as the tendency of two or more words that co-occur in discourse which are lexically or syntactically fixed to a certain degree (Schmitt, 2000; Nesselhauf, 2005), while colligation refers to the inter-relationships of words and grammatical items or the grammatical company a word keeps and the position it prefers (Firth, 1957; Hoey, 2000).

Specifically, this work addresses two central questions:

1. What are the collocational patterns found among the selected semi-technical words? and
2. What are the colligational patterns found among the selected semi-technical words?

METHODOLOGY

The methodological base of a corpus research is diverse, as it not only covers the fields of

corpus linguistics but also involves content and discourse analysis in the process of analyzing the lexical and grammatical relationships of words in the text; therefore, it is a combination of quantitative and qualitative analysis. This study examines the science corpus through discourse analysis specifically looking into meaning and form and the interconnection of lexis and grammar through keyword analysis and collocational patterns. The relation between text and context and the interdependence and interrelationships between lexis and grammar is seen in the analysis of collocations, in which words are partially defined or identified by the other words that surround them (Lewis, 1993; Sinclair, 1991).

This study is concerned with data of language used in the prescribed upper secondary textbooks, specifically Form Four and Form Five (upper secondary), and uses corpora to investigate the language of Science.

All the prescribed science textbooks from both the form 4 and form 5 levels used throughout Malaysia were initially edited manually, deleting numbers, formulae and repeated rubrics and sub-headings using a liquid corrector. These texts were then scanned and converted into txt files which were later analysed using the WordSmith version 4.0 concordance software. Each science subject corpus comprised of three textbooks. Therefore, the main Science corpus (combination of the General Science, Biology, Chemistry, and Physics textbooks) comprised of 12 textbooks.

Wordlists of each subject and a wordlist for the entire Science corpus were created. Table I displays the composition of each sub-corpus and the main Science corpus.

As the main method of examining words in this work is through keyword analysis using the WordSmith Tools 4.0, there is a need to have a suitable reference corpus. This keyword function provides a glimpse of what the text is about as the list is based on unique words that are frequent in the text (Reppen, 2001). WordSmith tools finds keywords by first generating frequency sorted word lists for the reference corpus and then for the research text/s. Each word in the research text is then compared with its equivalent in the reference text and the programme decides whether there is a statistically significant difference between the frequencies of the word in the different corpora. The requirement for a word list to be accepted as reference corpus by the WordSmith tools software is that it must be larger than the study corpus.

For a reference corpus to be selected, it should be five times larger than the study corpus (Berber-Sardinha, 2006). As the main Science corpus in this study has half a million words, a reference corpus of at least 2.5 million had to be selected. As there was no readily available corpus of about 2.5 million words on the English language used by Malaysians, the researchers decided on the 100 million word British National Corpus (BNC) as it is an established and reliable (Scott, 2001: 2002) corpus. As the English used in Malaysia leans more towards British

TABLE 1
Composition of the Science corpora

	Overall file size	Overall tokens (Running words) In text	Overall types (Distinct words)
Main Science Corpus	3,837,525	583,600	14,773
Biology	1,254,926	189,066	8,661
Chemistry	855,958	127,957	5,669
Physics	734,136	115,194	5,583
General Science	992,505	151,383	8,977

TABLE 2
Distribution of technical words in the Science corpus

Total no. of keywords	Non-technical	Sub-technical	Semi-technical	Technical
3113	40%	20%	9%	31%

English, a corpus focusing on British English was sought. The decision to use the BNC as the reference corpus for this study was also based on the procedure advocated and adopted by other analysts (Johnson, Culpeper and Suhr, 2003; Scott, 2000: 2001: 2002; Tribble, 2000). The BNC word-list used in this study was constructed by Scott and downloaded from his web page (http://www.lexically.net/wordsmith/).

Each word in the Science corpus is compared with its equivalent in the reference corpus and then the programme (WordSmith tools) decides whether there is a statistically significant difference between the frequencies of the word in the different corpora by evaluating the difference between counts per token and the total number of words in each text. The keyness of a word in this work is based on the log likelihood stats (Scott, 1996, WordSmith keywords Help file). The wordlist is then re-ordered in terms of the keyness of each word. 3113 positive keywords (unusually frequent words in the study corpus in comparison to the reference corpus) were identified. Only positive keywords were extracted as these were keywords which are more unique to the Science corpus.

Based on Cowan's (1974) and Nation's (2001) technical vocabulary categories, coding was then carried out by one of the researchers and an independent coder. The independent coder was chosen based on the criteria that the coder has had experience teaching EST for the past 5 years and was familiar with scientific words and general English language. The coder had also to agree to undergo coding training sessions with the researcher. The researcher involved in the coding was familiar with both scientific English and general English language.

Both the researcher and the independent coder were given a summary of Nation's, Cowan's and Godman and Payne's description of technical and sub-technical vocabulary to be read and clarified if it contained ambiguous statements. Then, coding of the keyword lists of technical words was carried out in three sessions. Coding differences were discussed and clarified, using concordancing lines of the text concerned and with reference to the *Oxford Advanced Learners' Dictionary* (2005), the *Collins Cobuild Dictionary* (2006), and the *Oxford Dictionary of Science* (2005).

Cohen's Kappa was used to assess inter-rater reliability. The inter-rater reliability was found to be high at 0.83, at $p<0.001$ with a 95% confidence interval (Cohen, as cited by Orwin, 1994). A 94% agreement between the two coders was also registered. Table 2 presents the distribution of the range of technical words in the Science corpus keyword list.

Once the semi-technical words were identified, the researchers checked each keyword against the individual subject wordlists to extract similar semi-technical words used frequently across the science subjects. Table 3 below presents the list of the similar semi-technical keywords according to keyness in the main Science corpus (combination of all science subjects). The table also displays the frequency in the Science corpus which these keywords are key and the percentage of the frequency of these keywords in the Science corpus against the number of running words in the Science corpus (583,600).

To analyse the behaviour and patterns of the semi-technical vocabulary, the four most key semi-technical words, 'reaction', 'cell', 'pressure' and ' mass', are examined in detail. Immediate 2-word collocations and frequent clusters of these words were extracted and analysed.

TABLE 3
Similar semi-technical keywords used in all four science subjects

Word	Frequency in Science corpus	Freq. % in Science corpus	Keyness
Reaction	1,449	0.25	7,890.78
Cell	1,217	0.21	6,263,06
Pressure	855	0.15	2,675.71
Mass	785	0.13	3,030.36
Nucleus	393	0.07	2,660.92
Volume	434	0.07	1,447.07
Materials	335	0.06	827
Negative	261	0.04	692.92
Function	289	0.05	524.77
Constant	201	0.03	420.63
Positive	213	0.04	294.99
Terminal	96	0.02	246.83
Acts	141	0.02	234.28
Fibre	81	0.01	206.92
Variables	73	0.01	133.33

RESULTS

Reaction

There were common syntactic characteristics found of the immediate 2-word collocations of this word. The collocations consisted of the combinations of 'noun+noun', 'adjective+noun', and 'noun+verb'. The position of the keyword 'reaction' in the 'noun+noun' combinations show that this keyword appeared both in the head and base positions of the collocations, showing the commutability (Cowie, 1981) of this keyword (*Fig. 1*). *Fig. 2* presents a cross section of the concordance lines of 'reaction' in the main science corpus.

The collocations with the word 'reaction' in the base positions – 'displacement reaction', 'chain reaction', 'redox reaction', show a semantic prosody associated with type of processes. Many of these combinations are fixed structures which form common highly technical compound nouns used in the scientific field (*Oxford Science Dictionary*). Two prosodic groups associated with 'type' and 'degree' were found among the 'adjective+noun' collocational combinations (*Fig. 3*).

Combinations such as 'chemical reaction', 'endothermic reaction', and 'exothermic reaction' referred to the prosodic group of type while combinations such as 'fast reaction', 'dark reaction' and 'light reaction' referred to the prosodic group of degree. The word 'reaction' was also seen to colligate frequently with verbs in the present tense (*Fig. 4*).

There was a strong colligational tendency of this word with a range of prepositions, as shown in Table 4 below.

The combinations of the word 'reaction' and the prepositions 'of', 'in', and 'between' seem to be common lexical patterns in this scientific discourse.

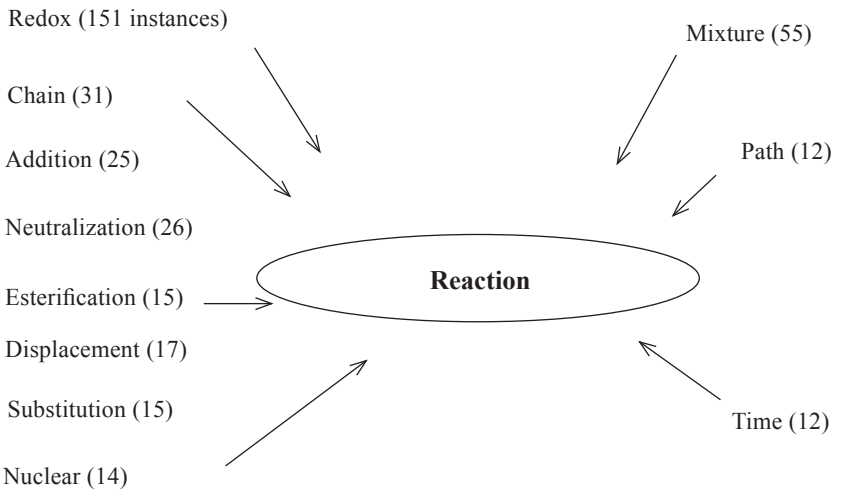

Fig. 1: Noun+noun combinations of the word 'reaction'

4	the leg to swing forward Calculating reaction time Rest the elbow on a table
5	Use the following equation to calculate reaction time Distance Reaction time
52	reactant product how the chemical reaction progresses as time passes
53	language support The chemical reaction progresses means the chemical
54	progresses means the chemical reaction moves forward or onwards The
55	or onwards The progress of a chemical reaction Progress of reaction particle of
348	of fat molecules The esterification reaction between fatty acids and gylcerol
370	fuels Burning fuels involves a redox reaction An oxidizing agent oxidizes a
371	a substance It is reduced in the redox reaction A reducing agent reduces a
372	a substance It is oxidized in the redox reaction Oxidation and reduction in

Fig. 2: Cross-section of the main Science corpus concordance lines of the word 'reaction'

Cell

There were similar lexical patterns shared among the collocations of 'reaction' and 'cell'. The syntactic characteristics of the collocations of the word 'cell' were similar to that of the word 'reaction' except with fewer 'noun+verb' combinations (shown in *Figs. 5-8*). *Fig. 5* presents a cross section of the concordance lines of 'cell' in the main science corpus.

Similar to the semantic prosody groups found in the collocations of the word 'reaction', the prosodic group associated with 'type' was found among the noun+noun and adjective+noun combinations – 'plant cell', 'animal cell', 'sickle cell' and 'dry cell'.

These collocations could be assumed to be free collocation combinations, if it based on the criterion of commutability (Cowie, 1981), as the word 'cell' is seen to collocate with a variety of nouns both in the head and base positions of the collocations (as seen in *Fig. 4*). For example, the combination 'cell+body' and 'blood+cell' can be seen as a free combination as 'body' also

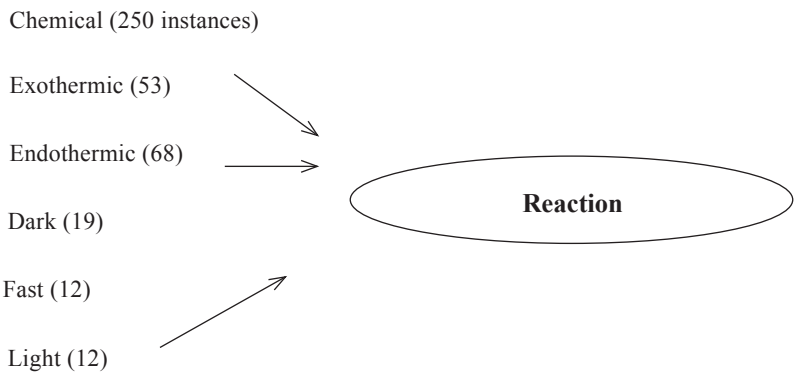

Fig. 3: Adjective+noun combinations of the word 'reaction'

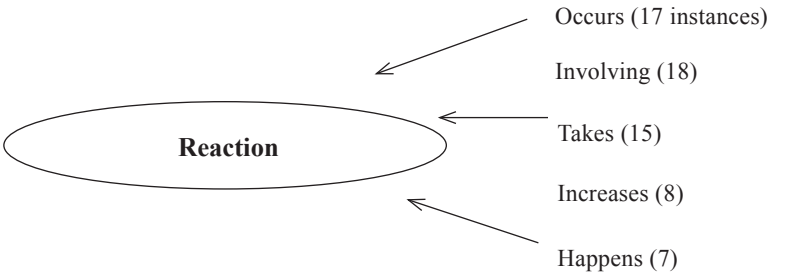

Fig. 4: Noun+verb combinations of the word 'reaction'

collocates with a variety of nouns as in 'human body', 'body pain' and 'blood' collocates with other nouns to form collocations such as 'blood group' and 'animal blood'. Therefore, none of the collocations seem to be restrictive as in the position or use of word; many words seem to fit in any one of those positions.

However, each element in the collocation does not necessarily carry a literal meaning. In fact, many of the elements in the collocations tend to acquire extended meanings. For example, in the collocations 'cell body' and 'blood cell', the elements in both collocations retain its individual meaning, thus, in compound form do not pose a problem in inferring meaning. However, the collocations 'companion cell' and 'guard cell' are difficult to define or to decode (Master, 2003), as the premodifiers 'companion' and 'guard' do not completely retain their original meanings of 'partner' and 'protector'. According to the *Oxford Dictionary of Science* (2005), the compound 'companion cell' is a type of cell found within the phloem of flowering plants which has a vague function of loading and unloading sugar whilst 'guard cell' refers to the stoma in leaves. Thus, most of the collocations found here seem to fall into the category between free and restricted combinations (Howarth, 1996). This distinction between literal and extended meanings can be further explained by

TABLE 4
Frequent prepositions with the word 'reaction'

Preposition	Head position	Word	Base position
Of	417	Reaction	97
In	10	Reaction	124
Between		Reaction	100
At		Reaction	47
To		Reaction	41
With		Reaction	26
For	6	Reaction	19

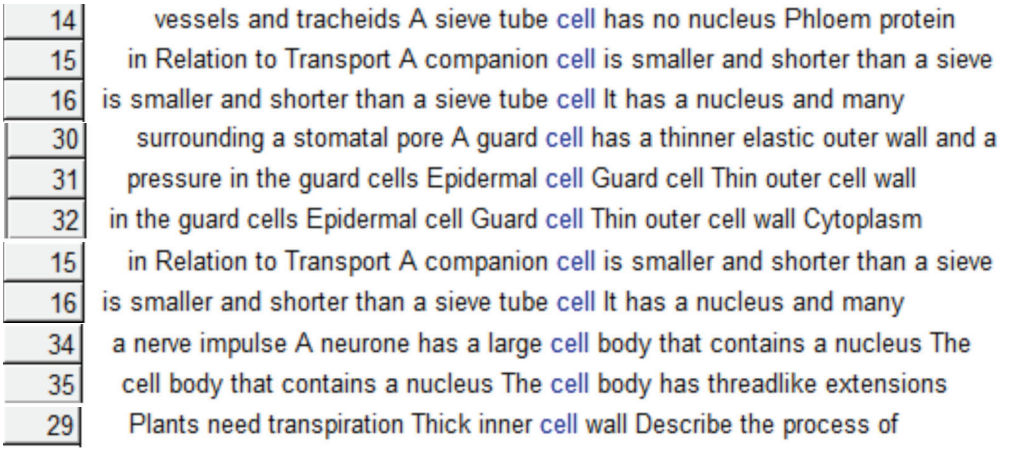

Fig. 5: Cross-section of the main Science corpus concordance lines of the word cell

looking at the various collocations of the word 'cell' which also formed compound nouns.

Many of the collocational combinations which formed compound nouns with similar syntactic characteristics or lexical patterns (noun+noun and adjective+noun) did not have similar semantic associations. For example, the compounds 'dry cell', 'voltaic cell', and 'chemical cell', all have similar syntactic characteristics of 'adjective+noun' with the keyword 'cell' holding the base position in all three combinations. A learner could assume or as Hoey (2007) states, learners could prime that these combinations are a type of cell or have the properties of 'chemical', and 'dry'. However, this priming would be inaccurate as 'dry cell' is a type of cell (battery), 'voltaic cell' is a device and 'chemical cell' relates to the chemical reaction of the cell (*Oxford Dictionary of Science*, 2005).

In the compounds 'cell body', 'cell walls', and 'cell membrane', the keyword 'cell' maintains the head position thus behaving as a premodifier of the base word. If general English language grammar rules are used to infer the meanings of these compounds, they could be understood as a type of body or type of wall. However, this transfer of grammar rules cannot be applied. Only the compound 'cell membrane' is a type of membrane as 'cell' modifies 'membrane'. Nonetheless, in the compounds 'cell wall' and 'cell body', the words 'wall' and 'body' are postmodifiers to the word

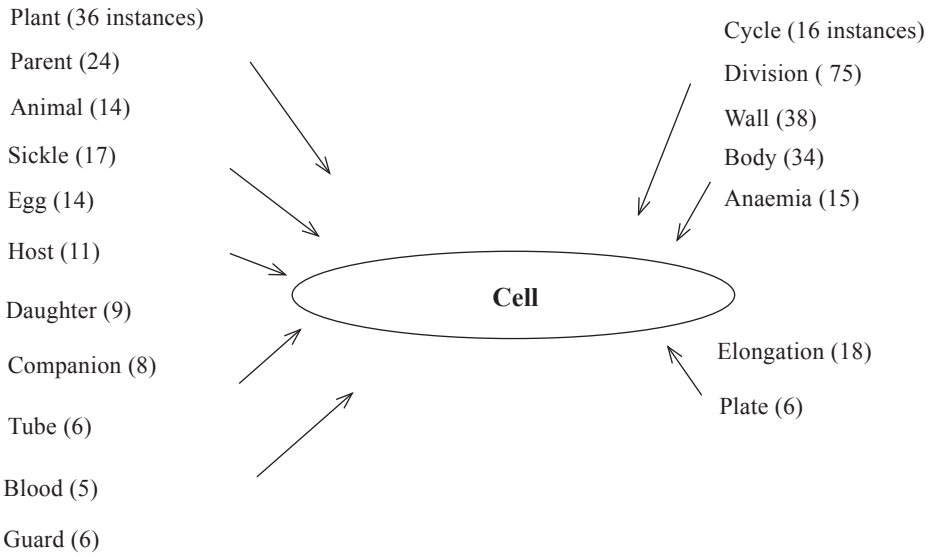

Fig. 6: Noun+noun combinations of the word 'cell'

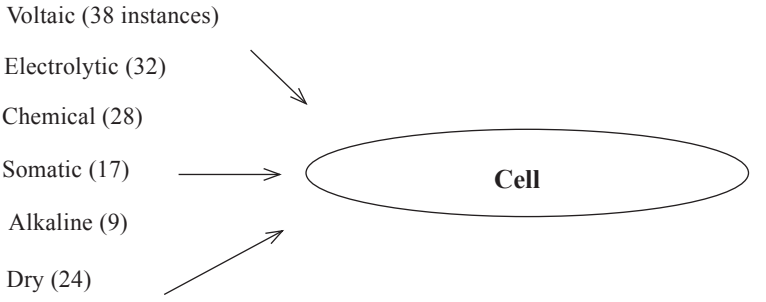

Fig. 7: Adjective+noun combinations of the word 'cell'

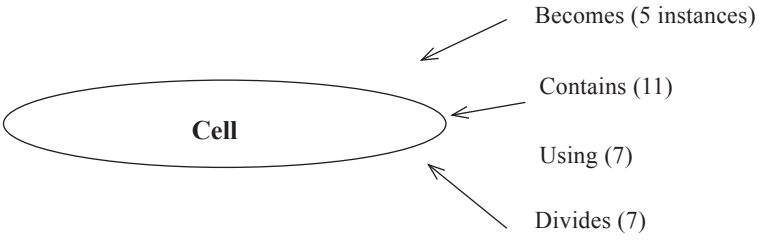

Fig. 8: Noun+verb combinations of the word 'cell'

TABLE 5
Frequent prepositions with the word 'cell'

Preposition	Head position	Word	Base position
Of	55	Cell	13
In	12	Cell	23
To		Cell	31
With		Cell	12
From	5	Cell	6
During	10	Cell	
By		Cell	7
For	6	Cell	

'cell', thus bringing about the meaning, 'a part of the cell' – the wall of the cell, the body of the cell.

In another more technical and complex example, the compounds 'daughter cell' and 'parent cell' have the syntactic combination of 'noun+noun' and the premodifiers- 'parent' and 'daughter', come from the same semantic family. Both could be assumed to be types of cell, which is not incorrect, but by definition (*Oxford Dictionary of Science*, 2005), 'parent cell' refers more to the use or purpose of the cell rather than the type of cell while 'daughter cell' is a type of cell which exists due to a process or reaction. Once again general English language grammar rules cannot be applied directly to the language used in Science. Similar to the colligational patterns of the word 'reaction', the word 'cell' was also seen to colligate (though not as frequent) with verbs in the present tense – 'becomes', 'using' and 'divides'. There were colligational tendencies with a range of prepositions, as displayed in Table 5 below.

Pressure

Once again, there were repeated similar lexical patterns among the collocations with the syntactic characteristics of 'noun+noun', 'adjective+noun' and 'noun+verb' as presented in *Fig. 10* below. However, there were less variety of collocations compared to the collocations formed by the words 'reaction' and 'cell'. *Fig. 9* presents a cross section of the concordance lines of 'pressure' in the main Science corpus.

Two semantic prosodies found among the collocations were the prosodies, type, and degree. As observed before, the semantic prosody associated with type was found among the 'noun+noun' and 'adjective+noun' combinations whilst the prosody associated with degree was found among the 'adjective+noun' combinations.

The collocations formed by the word 'pressure' seem to be more restricted compared to the collocations of the previous two keywords, as most of the nouns in these collocations seem to be in the head position followed by the keyword 'pressure'. There was only one frequent combination 'pressure cooker' with the keyword in the head position. The verbs which colligated with this word were both in the present and past tenses with most of them being lexicalized verbs or verbs which carry contextual meanings. There was colligational tendencies with prepositions but with a smaller set of prepositions such as 'of' (163 instances), 'in' (19 instances), 'on' (23 instances), and 'to' (19 instances). Only the prepositions 'of' and 'in' appeared in both the head and base positions of the collocations.

12	when the blood reaches the veins the pressure produced by the heart is
13	and press on the veins The blood pressure increases forces open the
14	to beat faster An increase in the partial pressure of carbon dioxide in the blood
15	it important to maintain a normal blood pressure level It should be less than
53	relaxes and vice versa Fast air flow Low pressure aerofoil Slow air High pressure
54	flow Low pressure aerofoil Slow air High pressure Wing as an aerofoil What will
59	body temperature and the osmotic pressure of blood To study human and

Fig. 9: Cross-section of the main Science corpus concordance lines of the word pressure'

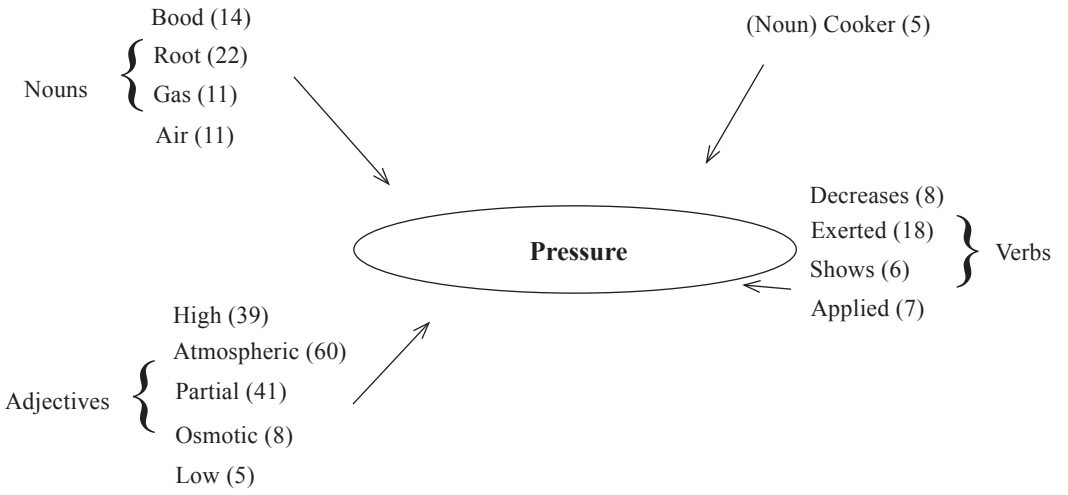

Fig. 10: Various syntactic combinations of the word 'pressure'

Mass

The common shared lexical patterns of 'noun+noun' and 'adjective+noun' were also found in this set of collocations (*Fig. 12*). However, there were no frequent 'noun+verb' combinations computed. *Fig. 11* presents a cross section of the concordance lines of 'mass' in the main Science corpus.

Only one semantic prosody associated with type was found among the 'noun+noun' and 'adjective+noun' combinations – 'body mass', 'bone mass', 'atomic mass', 'molecular mass'. Some of the collocations formed could be assumed to restricted collocations which are genre specific as the elements in the collocations do not carry a literal meaning but acquire extended meanings which are technical in nature. For example, the collocations 'mass number', 'critical mass', and 'atomic mass' are difficult to decode as the postmodifier 'number' and premodifiers 'critical' and 'atomic' do not completely retain its general English language meanings. The compound 'mass number' refers to a measurement specifically the number of nucleons in an atomic nucleus of a particular nuclide, 'critical mass' refers to the minimum mass of fissile material and 'atomic mass' refers to a unit of measurement (*Oxford Dictionary of Science*, 2005).

1	is a disease in which bone mass is reduced and the bones become
2	and jogging will also increase bone mass We are unaware of how the
20	of accuracy desired Explain what dry mass is Growth Curves Growth curves
21	such as length height fresh mass or dry mass against time The
116	Nuclear Energy of the atomic mass unit or is mass of the carbon atom
123	this oxidation is transmitted to a known mass of water and the corresponding
124	for food leads to a severe loss of body mass Anorexics are ten times more

Fig. 11: Cross-section of the main Science corpus concordance lines of the word mass'

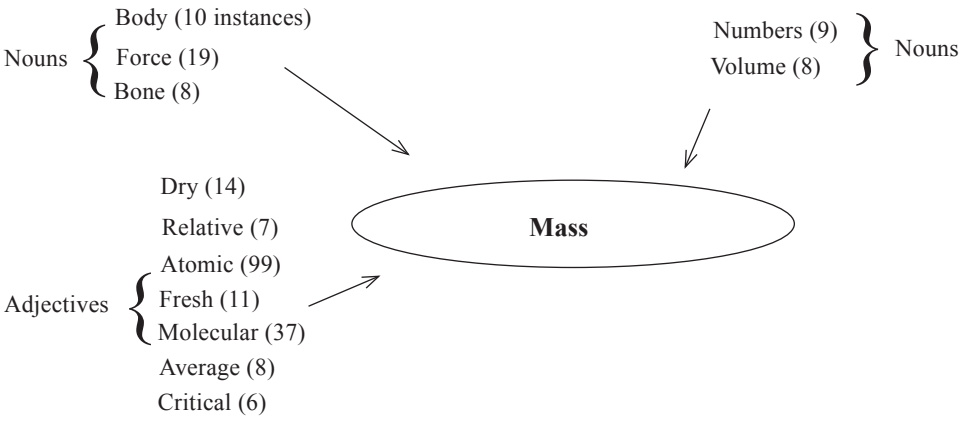

Fig. 12: Various syntactic combinations of the word 'mass'

What should be noted is that many non-technical words such as 'number', 'critical', 'fresh', 'dry' have turned into more technical words with extended meanings attached to them when in compound form. These compounds now acquire extended meanings more specific to the scientific field and are more technical in hierarchy. This has enormous pedagogical implications as the core subject and EST teachers should explicitly teach these words in context to show the variation in use and meaning.

DISCUSSION

Although the percentage of semi-technical vocabulary found in this Science corpus was small (9%), this is a known problematic area (Herbert, 1965; Trimble, 1985) which needs to be focused upon in the classrooms. This work supports Trimble's (1985) conclusion on semi-technical words being confusing and containing layers of complexity which would pose a problem to second language learners learning science in English.

The analysis of the semi-technical words generated many chunks and collocations which are found to be common and significantly used in the prescribed science textbooks. There are common lexical and grammatical patterns, the most common being the 'noun+noun' and the 'adjective+noun' combinations. The combinations with the same grammatical pattern seem to share aspects of meaning. This study found the semantic prosody of type was realized by the grammatical patterns of 'noun+noun' and 'adjective+noun' whilst the semantic prosody of degree was realized by the 'adjective+noun' pattern. These patterns could be taught to

students so as to prepare them for possible collocations, once the prosodic group has been identified. The pedagogical implication is that some sense and grammar patterns can be focused upon more than others.

The analysis on the four semi-technical words also show that the collocations of these words very often form compound nouns and multi-word units which do not retain the literal meanings of the elements in the combinations, but rather acquire extended meanings which were more technical and genre specific. The observation by Trimble (1985) and Thirumalai (2003) that the language of science is riddled with multi-word units and compound nouns which are complex together with the findings in this study, reinforce the importance of identifying collocational patterns in a science corpus.

Students need to understand why general English language rules cannot be applied all the time to infer meanings of compounds such as in the compounds 'cell body', 'cell wall', and 'cell membrane'. They need to understand the different meanings of these compounds and why even though they share similar syntactic characteristics, the flexibility of some combinations is arbitrarily blocked by usage, thus becoming genre specific collocations.

Currently, the exposure given to the upper secondary students in Malaysia on compound nouns is via the prescribed English for Science and Technology (EST) textbooks. The current EST textbooks deal with compound nouns only at the introductory level with exercises requesting students to either identify compound nouns or define them through pair exercises or form compound nouns either by inserting the 'ing' form and affixes. None of these exercises are able to explain the complexity of compound nouns in science and explain the extended meanings acquired by many of them.

CONCLUSIONS

This study has shown that scientific English lexis and phraseology, especially the ones related to semi-technical vocabulary are significantly different from the general English language as it contains terminology with limited meanings and words which acquire extended meanings. General English language grammar rules most often cannot be used to infer the meanings of compound nouns in Science texts as many of the elements in the 2-word combinations do not retain their literal meaning but instead acquire extended meanings more specific to the scientific field. However, the phraseology and patterns identified should not be over-represented. Students should be informed of the typical patterns but be reminded of the diversity and flexibility of these patterns in scientific discourse.

An effective way for learners to increase their active vocabulary is for them to be centrally involved in the learning process (Tognini-Bonelli, 2001; Nation, 2001). In the learning and acquisition of specialized or technical vocabulary it is important to develop vocabulary in a systematic way rather than by incidental learning. Through exposure to the collocational and colligational patterns of words, learners will be able to understand how words are formed and the type of structures they appear in. This knowledge will then allow them to develop their own strategies for inferring meanings from context.

The analysis on the four semi-technical words showed that these words were not used similarly and with the same meaning across the four Science subjects. The words were often used with extended meanings attached (not as used in general English language contexts) to them, especially in compound form. This proves the enormity of learning semi-technical words and importance of understanding the variation in the meaning of the words. These types of words should be focused upon and taught or given more exposure by the EST and content teachers so that students are aware of the differences in meaning and use of the words. This study has shown the importance of being aware of the variation in language especially in specialized and academic texts. Corpus-based research has allowed us to

answer many questions concerning language formation. Instead of telling learners 'it depends' on the context or sentence, corpus driven data can show 'what it depends on'.

REFERENCES

Ansary, H. and Babaii, E. (2003). A 'Bell-Jar' consensus-reached set of universal characteristics of ELT textbooks. In J. Mukundan (Ed.), *Readings on ELT materials* (pp. 69-83). Serdang: Universiti Putra Malaysia Press.

Ary, D., Jacobs, L.C. and Razavieh, A. (1985). *Introduction to Research in Education* (3rd Edn.). New York: Holt, Rhinehart & Winston.

Atkinson, D. (1999). *Language and Science Annual Review of Applied Linguistics*, *19*, 193-214. Retrieved on July 12, 2006 from http://www.uefap.com/writing/research/langsci.htm.

Barber, C.L. (1962). Some measurable characteristics of modern scientific prose. In J.M. Swales (Ed.), *Contributions to English syntax and philology*. (1985). *Episodes in ESP* (pp. 3-16). Oxford: Pergamon Press Ltd.

Berber-Sardinha, T. (2006). Comparing corpora with WordSmith Tools: How large must the reference corpus be? Retrieved on October 14, 2007 from http://citeseer.ist.psu.edu/cis?q=tony+Berber+Sardinha.

Biber, D., Conrad, S. and Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Burrough-Boenisch, J. (2003). Examining present tense conventions in scientific writing in the light of reader reactions to three Dutch-authored discussions. *English for Specific Purposes*, *22*, 5-24.

Chiapetta, E. L., Sethna, G.H. and Fillman, D.A. (1991). A quantitative analysis of high school Chemistry textbooks for scientific literacy themes and expository learning aids. *Journal of Research in Science Teaching*, *28*, 939-951.

Collins COBUILD Advanced Learner's English Dictionary (5th Edn.). (2006). Glasgow: HarperCollins Publishers.

Cowan, J.R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, *8*(4), 389-399.

Cowie, A.P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, *2*, 223-235.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age questions and some other matters. *Working Papers on Bilingualism*, *19*, 121-129.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada. A reassessment. *Applied Linguistics*, *2*, 132-149.

Firth, J.R. (1957). *Papers in Linguistics, 1934-1951*. London: Oxford University Press.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ Online Journal*, March *8*(4). Retrieved on June 16, 2007 from http://tesl-ej.org/ej32/al.html.

Gavioli, L. (1997). Exploring texts through the concordancer: Guiding the learner. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and language corpora* (pp. 83-89). New York: Addison Wesley Longman.

Gledhill, C. (1996). The phraseology of rhetoric, collocations and discourse in cancer abstracts. Languages for Specific/Academic Purposes. Retrieved on July 20, 2007 from http://ec.hku.hk/kd96proc/authors/papers/gledhill.htm.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115-135.

Gottfried, S.S. and Kyle, W.C. (1992). Textbook use and the Biology education desired state. *Journal of Research in Science Teaching*, *29*, 33-49.

Halliday, M.A.K. and Martin, J.R. (1993). *Writing Science. Literacy and Discursive Power*. London: The Palmer Press.

Herbert, A.J. (1965). The structure of technical English. In J. Swales (Ed.), *Episodes in ESP*, 17-27. Oxford: Pergamon Press Ltd.

Higgins, J.J. (1967). Hard facts. (Notes on teaching English to science students). Reproduced in J. Swales (Ed.), (1985). *Episodes in ESP*, (pp. 28-37). Oxford: Pergamon Press Ltd.

Hoey, M. (2000). A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocations,* 224-242. Hove: Language Teaching Publications.

Hoey, M. (2007). Lexical priming and literacy creativity. In M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis,* 7-29. London/ New York: Wolfgang Continuum.

Howarth, P. (1996). *Phraseology in English Academic Writing. Some Implications for Language Learning and Dictionary Making*. Tubingen: Niemeyer.

Hunston, S. and Francis, G. (2000). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Johnson, S., Culpeper, J. and Suhr, S. (2003). From 'politically correct councillors' to 'Blairite nonsense': Discourses of 'political correctness' in three British newspapers. *Discourse and Society*, *14*(1), 29-47.

Kornwipa, P., Somcheon, H. J. and Cowan, R.A. (2001). The teaching of academic vocabulary to Science students at Thai University. *Studies in Languages and Language Teaching*, *10*, December, 51-64. Retrieved July 10, 2006 from http://www.sc.mahidol.ac.th/sclg/sllt/html/year-2001.html.

Krashen, S. and Biber, D. (1988). *On Course*: *Bilingual Education's Success in California*. Sacramento: California Association for Bilingual Education.

Lackstrom, J., Selinker, J. and Trimble, L. (1975). Grammar and Technical English. The art of TESOL, part 2. *English Teaching Forum*, *XIII* (3 and 4), 250-260.

Laplante, B. (1997). Teaching Science to Language Minority Students in Elementary Classrooms. *NYSABE Journal*, *12*, 62-83.

Lewis, M. (1993). *The Lexical Approach*. Hove, England: LTP.

Marco, M.J.L. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes*, *19*, 63-86.

Master, P. (2003, July). Noun compounds and compressed definitions. *English Teaching Forum.* Retrieved on July 8, 2006 from http://iteslj.org/links/TESL/Articles/Grammar.

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

Nelson, M. (2001). A corpus based study of business English and business English teaching materials. Unpublished PhD Thesis. Manchester: University of Manchester.

Nesselhauf, Nadja. (2005). *Collocations in a learner corpus*. Philadelphia: John Benjamins Orwin, R.G. 1994. Evaluating coding decisions. In H. Cooper and L. Hedges (Eds.), *The handbook of research synthesis* (pp.139-162). New York: Russell Sage Foundation.

Oxford Advanced Learner's Dictionary (seventh edition). (2005). Sally Wehmeier (Ed). New York: Oxford University Press.

Oxford Dictionary of Science (fifth edition). (2005). New York: Oxford University Press.

Reppen, Randi. (2001). Review of MonoConc Pro and WordSmith Tools. *Language Learning and Technology*, *5*(3), May, 32-36. Retrieved on November 12, 2006 from http://llt.msu.edu/vol5num3/review4/.

Rosenthal, J.W. (1996). *Teaching Science to Language Minority Students*. England: Multilingual Matters Ltd.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Scott, M. (1996, 1997, 1999). *Versions (1.0, 2.0, 3.0, 4.0) WordSmith Tools*. Oxford: Oxford University Press.

Scott, M. (2000). Focusing on the text and its key words. In L. Burnard and T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 103-122). Frankfurt, Germany: Peter Lang.

Scott, M. (2001). Comparing corpora and identifying key words, collocations and frequency distributions through the WordSmith Tools suite of computer programs. In M. Ghadessy, A. Henry and R.L. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 47-67). Amsterdam/Philadelphia: John Benjamins Publishing Co.

Scott, M. (2002). Picturing the keywords of a very large corpus and their lexical upshots or getting at the Guardians' view of the world. In B. Kettemann and G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 43-50). Amsterdam: Rodopi.

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

Soler, V. (2002). Analysing adjectives in scientific discourse, an exploratory study with educational applications for Spanish speakers at advanced university level. *English for Specific Purpose*s, *21*(2), 145-165.

Spurlin, Q. (1995). Making science comprehension for language minority students. *Journal of Science Teacher Education*, *6*(2), 71-78.

Stern, H.H. (1992). *Issues and Options in Language Teaching*. Oxford: OUP.

Swales, J. (1998). *Other Floors, Other Voices: A Textography of a Small University Building*. Mahwah, NJ: L. Erlbaum.

Tarone, E., Dwyer, S., Gillette, S. and Icke, V. (1981). On the use of the passive in two astrophysics journal papers. Reproduced in John Swales, (Ed.), (1985) *Episodes in ESP* (pp.188-208). Oxford: Pergamon Press Ltd.

Thirumalai, M.S. (2003). Language in Science. *Electronic Journal of Language in Science*, *3*(1), January 2003. Retrieved on January 7, 2007 from http://www. languageinindia.com/jan2003/ languageinscience.html.

Thompson, D.R. and Rubenstein, R.N. (2000). Learning Mathematics vocabulary: Potential pitfalls and instructional strategies. *Mathematic Teacher*, *93*(7), 568-573.

Tognini-Bonelli, Elena. (2001). *Corpus Linguistics at Work*. Philadelphia: John Benjamins.

Tomlinson, B. (1998). Access-self materials. In Tomlinson (Ed.), *Materials development in language teaching* (pp. 320-337). Cambridge: Cambridge University Press.

Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard and T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75-90). Frankfurt, Germany: Peter Lang.

Trimbe, L. (1985). *English for Science and Technology: A Discourse Approach*. New York: Cambridge University Press.

Willis, J. (1998). Concordances in the classroom without a computer: Assembling and exploiting concordances of common words. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 44-67). Cambridge: Cambridge University Press.