International Education Studies; Vol. 5, No. 6; 2012
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Collocations of High Frequency Noun Keywords in Prescribed Science Textbooks

Sujatha Menon[1] & Jayakaran Mukundan[2]

[1] Academy of Language Studies, Universiti Teknologi MARA, City Campus Melaka, Malaysia

[2] Department of Language Education & Humanities, Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia

Correspondence: Sujatha Menon, Academy of Language Studies, Universiti Teknologi MARA, City Campus Melaka, 110, Off Jalan Hang Tuah, 75300 Melaka, Malaysia. Tel: 60-6283-3375/60-126-595-875. Email: sjathark@yahoo.com

Received: September 7, 2012 Accepted: September 24, 2012 Online Published: October 12, 2012

doi:10.5539/ies.v5n6p149 URL: http://dx.doi.org/10.5539/ies.v5n6p149

Abstract

This paper analyses the discourse of science through the study of collocational patterns of high frequency noun keywords in science textbooks used by upper secondary students in Malaysia. Research has shown that one of the areas of difficulty in science discourse concerns lexis, especially that of collocations. This paper describes a corpus-based analysis of the lexical collocations, specifically the noun/noun (NN), noun/adjective(AN) and verb/noun(VN) combinations, of high-frequency noun keywords in this science corpus. The study found that the NN and AN syntactic combinations were the more predominant combination type among the collocations of the noun keywords. Most of the collocations have been found to fall into the category between free and restricted combinations. In addition, the analysis have also uncovered clusters of overlapping collocations which show a tendency towards the free end of the restricted category of combinations.

Keywords: nouns, keywords, collocation, science, EAP

1. Introduction

1.1 Overview

This paper describes a corpus-based analysis of the lexical collocations, specifically the noun/noun (NN), noun/adjective (AN) and verb/noun (VN) combinations, of high-frequency noun keywords in a science corpus. This present study is part of a thesis investigating the language of science in prescribed textbooks used in Malaysian secondary schools. The main aim of the thesis was to investigate the patterns and phraseology of words used in the prescribed upper secondary science textbooks.

It has been widely accepted that the knowledge and use of collocations are a pre-requisite to proficient language use (Gledhill, 2000; Hill, 2000; Howarth, 1998; Sinclair, 1991; Wray, 1999, 2002). However, studies have shown that L2 learners face problems with collocations in their written and spoken language (Howarth, 1998; Nesselhauf, 2005) and that L2 learners rely heavily on language rules rather than lexicalixed routines and context appropriateness when forming language chunks (Foster, 2001; Skehan, 1998). This implies that learners would often use words individually without taking note of the environment or context and thus have a tendency to combine words that do not belong or 'mis-chunk'.

This problem is further aggravated when learners try to comprehend academic texts. Comprehending academic language is crucial as it is the language used in instruction, textbooks and exams. Academic language such as the language used in science, may not differ very much in structure and vocabulary from language used in daily social interactions (Durrant, 2009; Gledhill, 2000; Menon and Mukundan, 2010; Peacock, 2012; Trimble, 1985) but it usually contains vocabulary, grammatical patterns and phraseology particular to that discipline and discourse. Research through decades has shown that one of the areas of difficulty and confusion in science discourse concerns lexis, especially that of lexical collocations and sub-technical vocabulary (Gledhill, 1996, 2000; Herbert, 1965; Peacock, 2012; Trimble, 1985; Ward, 2007).

1.2 Corpus-based Research on Science Texts

In the past few decades, there have been many corpus-based studies on the language of science texts. However, most have been on the academic discourse in science and technical research articles (Durrant,2009; Gledhill, 1996,2000; Marco, 2000; Peacock,2012; Soler,2002; Swales, 1998; Tarone, Dwyer, Gillette and Icke, 1981; Williams,1998), while some carried out studies on textbook discourse and teaching and learning materials (Barber, 1962; Higgins, 1967; Lackstrom, Selinker and Trimble, 1975; Kornwipa, Somcheon and Cowan, 2001; Mudraya, 2006; Ward, 2009).

Much of the earlier published work have focused much on the analysis of specific grammatical features and verbs (Barber, 1962; Higgins, 1967), passive and active forms and tenses (Lackstrom, et al., 1975; Rodman, 1994, cited in Atkinson, p.197), hedging (Grabe and Kaplan, 1997, cited in Atkinson, 1999, p.6; Hyland, 1994) and on discourse functions (Rodman, 1991, cited in Atkinson, p.197; Swales, 1998). Collocational behavior was examined by a few, notably, Master(1991, cited in Atkinson, 1999, p.196) who looked at collocations of inanimate subjects with active-voice verbs in popular science newsletters and Gledhill (1996) who analysed collocations and phraseology of rhetorics in a corpus of cancer research articles.

Recent publications and studies on the language used in the science and technical subjects have thrown some light on collocations and multi-word units (Biber, Conrad and Cortes, 2004; Durrant, 2009; Gledhill, 2000; Marco, 2000; Mudraya, 2006; Peacock, 2012; Ward, 2007). However, most of these studies have been on collocational behavior observed in research articles, with the exception of Mudraya (2006) who looked at formulaic multi-word units or collocations in her Student Engineering English Corpus which consisted of 13 engineering texts and Ward (2007) who looked at specific nouns and their collocations in Chemical Engineering textbooks.

1.3 Collocations/Prefabricated Language

Collocation refers to the tendency of two or more words that co-occur in discourse which are lexically or syntactically fixed to a certain degree (Nesselhauf, 2005; Schmitt, 2000). They have been called prefabricated language (Pawley & Syder, 1983), phraseological units (Cowie, 1981), lexical phrases (Nattinger & DeCarrico, 1992; Nattinger, 1988), co-selection of words (Sinclair, 1996), multi-word units (Lewis, 2001; 1993), conventionalised forms, ready-made utterances (Wray, 2002) and word partnerships (Mudraya, 2006). Studies of language acquisition have shown that the acquisition of formulaic speech and chunks are central to the learning of language as these chunks could be stored and retrieved easily (Nattinger & De Carrico, 1992).

1.4 Importance of Collocations

Many linguists have listed the importance of studying collocations (Hill, 2000; Hunston & Francis, 2000; Pawley & Syder, 1983; Peters, 1983; Wray, 1999). Some of the more pertinent reasons:

1.4.1 Combinations are Predictable

The way words combine in collocations is fundamental in language use. Hill (2000) states that to an extent vocabulary choice is predictable and that some patterns can be learnt which would help learners in the acquisition of the language. This relates to Hoey's (2007) notion of lexical priming.

He states that every time a word is encountered by a learner, the learner subconsciously keeps a record of the context and co-text of the word. These encounters will prime the learners so that when the word is used later, it will be used with its typical collocations, grammatical function, in the semantic context in a familiar social context with similar pragmatics and to similar textual ends (2007, p.8).

This notion of the predictability of collocations is supported by Gledhill (2000). In his study of cancer research articles, he found that the use of collocations in scientific writing is mainly through knowledge of what are acceptable and appropriate conventions in scientific writing or knowing the best way of expressing certain ideas. This selection, he states, may not indicate that these are the best expressions to be used but is largely a feature of convention and acceptability within the discourse community.

1.4.2 Easily Retrieved

Collocations are easily retrieved (Hunston & Francis, 2000) and are usually stored and processed as unitary wholes (Schmitt and Carter, 2004). Collocation allows us to think and communicate more quickly and efficiently. The brain tends to process prefabricated units better (Hill, 2000; Pawley & Syder, 1983).

1.5 Approaches to Collocations

1.5.1 Frequency-based Approach

The first approach is the 'frequency-based approach in which a collocation is the co-occurrence of words at a certain distance based on frequency (Nesselhauf, 2004). Researchers adopting the frequency-based approach to collocations have differing views as to what constitutes collocations. Moon (1998) considers co-occurrences of all frequencies as collocations while Stubbs (1995) only accepts frequent co-occurrences. Others accept recurrent co-occurrences (more than once) in a given corpus as the defining criterion (Kennedy, 1990; Kjellmer, 1987 cited in Nesselhauf, 2005, p.13). In this approach, the syntactic relationship between the elements does not play a role in deciding whether they form a collocation or not (Nesselhauf, 2005).

1.5.2 Phraseological Approach

The second approach is the phraseological approach in which a collocation is seen as a type of word combination, most commonly as one that is fixed to some extent (Nesselhauf, 2004). This research adopts Cowie's (1994; 1981) phraseological approach to collocations. He divides word combination into two main types, 'composites' and 'formulae'. Formulae are combinations with a pragmatic function such as 'How do you do?' or 'Good morning' (Nesselhauf, 2005, p.14). Collocations, he adds, belong to the composites group with primarily a syntactic function. The distinctions in the group of composites are based on firstly, the transparency of the elements in a word combination, which refers to whether the elements in the combination has a literal or non-literal meaning, for example in the combination 'drink tea', both the elements are used in a literal sense while at least one of the elements in the combination 'perform a task' has a non-literal meaning (Nesselhauf, 2005, p.14). Secondly, it is based on the commutability or substitutability of the elements referring to the degree the substitution of the elements of the combinations are restricted (Nesselhauf, 2005).

On this basis, he distinguishes four types of combinations which are seen as forming a continuum (Nesselhauf, 2005, p.14):

1) Free combinations (eg. Drink tea)
All the elements of the word combination are used in a literal sense and the restrictions on substitution is based on semantic grounds.

2) Restricted collocations (eg. Perform a task)
Some substitution is possible and at least one element has a non-literal meaning.

3) Figurative idioms (eg. Do a u-turn (related to attitude or behaviour)
Substitution of elements is seldom possible. The combination has a figurative meaning but retains a literal interpretation.

4) Pure idioms (eg. Flying colours)
Substitution of the elements is impossible and the combination has a figurative meaning which does not have a current literal interpretation.

According to the phraseological approach (Cowie, 1981), there exists no relationship between the elements in a collocation. However, this work supports the views of Melcuk (1998) and Hausmann (1989) who state that there is a subtle if not direct relationship between the elements in a collocational combination. Melcuk (1998, cited in Nesselhauf, 2005, p.17) describes the element which has been selected on the basis of its meaning as the 'keyword' and the element the 'keyword' selects to express a certain meaning as the 'value'. This is similar to that expressed by Hausmann (1989, cited in Nesselhauf, 2005, p.17) who describes the element which is selected first in production as the 'base' and the element whose selection depends on the 'base' as the 'collocator'. The phraseological approach requires the elements of collocations to be syntactically related.

There are two types of collocations, lexical and grammatical. Collocations in which two lexical elements co-occur are lexical collocations, such as in any combination with nouns, verbs, adjectives and adverbs. Collocations in which a lexical and a grammatical element such as a preposition, infinitive or a clause co-occur are grammatical collocations.

1.6 Research Questions

This study is part of a larger project concerned with data of language used in textbooks for the Science subjects in form four and form five in Malaysia. The present study investigated two-word collocations of the most positive noun keywords in this science corpus. The researcher looked at Keywords as these are words whose frequency is unusually high in the study corpus in comparison to a reference text and thus keywords tend to represent those words which are characteristic of a discourse or a genre (Scott, 2001).

The following research questions were formulated:

1) What are the immediate two-word collocations of the most positive noun keywords in the science corpus?

2) What are the syntactic characteristics of these collocations?

2. Methodology

2.1 Selection of Texts for Corpus

Though there is heavy reservation against using textbook language as corpus data, as the criticism leveled against it is that it is not naturally occurring language, it should be noted that in second language learning contexts such as that in Malaysia, the students main exposure to English language is through formal education and school textbooks. Thus the texts used in this Science corpus consisted of the form five General Science, Physics, Biology and Chemistry textbooks from two zones (total of eight books) and the form four General Science, Physics, Chemistry and Biology textbooks (total of 4 books). There were 12 textbooks used to create the Science corpus. As the total number of running words in each textbook ranged between 37,000 to 69,000 words, the Science corpus consisted of 583, 600 words.

As this study examines a half a million word Science corpus, the reference corpus should be about 2.5 million words or more (Berber-Sardinha, 2006). As there were no readily available corpus of about 2.5 million words on the English language used by Malaysians, the researcher selected the 100 million word British National Corpus (BNC) as it is an established and reliable corpus (Scott, 2001; 2002). In addition, as the English language used in Malaysia leans more towards British English, a corpus focusing on British English was sought. This decision to use the BNC was also decided upon based on the procedure advocated and adopted by other analysts (Johnson, Culpeper & Suhr, 2003; Scott, 2000, 2001, 2002; Tribble, 2000). Specifically a word-list based on the entire BNC set of written texts and constructed by Scott, which is readily made available via his web page (http://www.lexically.net/wordsmith/) was utilised in this study.

2.2 Selection of Keywords

Once the texts were collected and all numbers and formulae were deleted manually, they were then scanned and converted into txt files to be analysed using the WordSmith version 4.0 concordance software. Wordlists and keyword lists were then created and word class categories were assigned by independent coders (the interrater reliability for the coders was found to be good at K= 0.68 with a 90% agreement). The keyword list of the science corpus was then reordered in terms of 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.

As there were too many keywords and analysis of all these words were not feasible, the researcher adopted the approach used by other analysts (Johnson & Ensslin, 2006; Nelson, 2001; Nishina, 2007) who analyzed a small sample of keywords. Thus, the researcher chose a sample of (confidence level of 95% and a 5% margin of error) 250 most frequent positive keywords from the Science corpus. As the distribution of nouns, verbs and adjectives in the entire Science corpus averaged around 67% nouns, 16% verbs and 15% adjectives, the researcher attempted to choose words according to the above percentage division, thus 170 noun keywords were selected.

As the purpose of this section is to identify similar or different lexis and grammatical structures which would be pedagogically useful, the above set of noun keywords were ‘loosely’ (as some of these words can belong to more than one semantic category) assigned semantic categories. However, as many of these terms were used in different contexts and subject areas with no clear demarcation of semantic boundaries, some of the words were not assigned semantic categories and thus, identified only as ‘scientific terms’ or ‘general terms’ used in Science. The semantic categories chosen were based on the explanation and definition provided in the Oxford Dictionary of Science (2005). Assigning the collocations of these keywords, grammatical and semantic categories would make this analysis more pragmatic and directly transferable to the classrooms. This study’s approach to complementation patterns of collocations follow Hunston and Francis’s (2000) approach of assigning formal components of a pattern (verb+noun, adjective+noun, noun+verb, etc) to a combination rather than carrying out structural interpretation (verb + object or verb + complement) of that combination. Table 1 presents the semantic categories assigned to the selected noun keywords.

Table 1. Semantic categories of the noun keywords

Semantic category	Keywords
Process	Reaction, combustion, oxidation, electrolysis, fertilization, fission, formation, meiosis, mitosis, neutralization, photosynthesis, radiation, reactivity, respiration, transpiration
Measurable substances	Temperature, heat, energy, pressure, velocity, momentum, resistance, speed, volume, voltage, wavelength
Related to forms of life	Cell, bacteria, body, organisms, microorganisms, animal, human, plants, fungi, palm, xylem, phloem, roots Blood, fats, brain, glucose, diseases, alleles, capillaries, chromosomes, gametes, gland, hormone, membrane, muscles, nerve, neurone, ovary, ovum, pathogens,, plasma, tissues, zygote
Elements on earth	Gas, water, air, carbon, hydrogen, chlorine, nitrogen, oxygen
Related to light, electricity and signals	Light, waves, electricity, rays, current, anode, bulb, cathode, circuit, coil, conductor, impulses, transformer, transmission,
Things/ Objects/ Substances	Metal, earth, oil, atom, solution, fuel, compounds, acid, additives, alcohol, alkali, alkanes, ammonia, auxins, chloride, diagram, electrons, enzymes, esters, ethanol, aluminium, graph, hydroxide, internet, ions, iron, latex, lens, magnesium, medicines, molecules, neutrons, nucleon, nucleus, nutrients, piston, plastics, polymers, potassium, proteins, proton, bromine, reactant, salts, soap, sodium, sulphate, sulphur, copper, tube, vessels, zinc, alkenes, food
General Terms	Substances, particles, mass, growth, force, object, catalyst, characteristics, fluid, collision, concentration, cycle, direction, displacement, effects, environment, equation, examples, experiment, flow, frequency, immunity, liquid, materials, moles, number, process, properties, rate, elements, surface, system, types, ecosystem,

The immediate two-word collocates for each word were computed using the concord-collocation function in WordSmith 4.0. This work only looked at two-word collocates as this type of collocation is the type which is focused upon in the vocabulary list provided in the syllabus outline. The minimum specification for collocations to be accepted was set as 5. Collocations containing pronouns, conjunctions, the verb ‘to be’ form and proper nouns were deleted, as these were not combinations in focus in this work.

3. Results

Analysis of Noun Keywords through Semantic Categories

3.1 Semantic Category: Process

The NN(noun+noun) and AN (adjective+noun) were the common type of combinations found in this semantic set.The NN combination were found to be the most common syntactic combination in the words ‘combustion’ (59%), ‘fission’ (77%) and ‘oxidation’ (90%). The position of the keywords in the NN syntactic combination showed that these nouns appeared both in the head and base positions of the collocates: ‘reaction rate’, ‘acid reaction’, ‘combustion chamber’, ‘stroke combustion’, ‘oxidation number’, ‘combustion oxidation’- showing the commutability of these keywords. Analysis of the positions of the keywords in the base positions of the combination (‘addition reaction’, ‘chain reaction’, ‘stroke combustion’, ‘complete combustion’, “bond formation’, ‘ovum formation’) show a semantic prosody associated with ‘type of processes’. Notable, was the frequency of some noun phrases; ‘redox reaction’ (151 instances), ‘chemical reaction’ (250 instances) and ‘oxidation number’ (146 instances). These collocates can be assumed to be restricted in its combination as they appear frequently in this combination across the Science corpus.

The AN combination formed the largest percentage of the syntactic combination in the words ‘reaction’ (53%), ‘radiation’ (86%) and ‘respiration’ (100%) thus, a majority of the words in combination with these three words acted as modifiers to the keyword such as in the combinations ‘radioactive radiation’ and ‘ anaerobic respiration’.

The NN and AN syntactic combinations show that the common grammatical patterns of collocations of nouns in Science texts are similar to that used in general English language (Hunston & Francis, 2000, pp56-58).

However, there was a notable lack of verb+noun (VN) collocations (3-6%) in this semantic category. Most of the verbs consisted of delexical verbs or verbs which do not carry any or much contextual meaning on their own such as ‘occurs’, ‘involving’, ‘happens’ and ‘takes’.

3.2 Semantic Category: Measurable Substances

The NN and AN combinations registered as the majority type of collocations only for three words in this category: ‘energy’ (50%), ‘voltage (69%) and ‘pressure’ (54%). The other words registered more instances of grammatical collocations which is not discussed in this paper. In the collocations of the words ‘velocity’, ‘momentum’, ‘resistance’, ‘speed’, and ‘voltage’, there was a high tendency for the words to collocate with an adjective rather than a noun.

The position occupied by the keyword in the NN combination differed from word to word. The words ‘temperature’, ‘resistance’ and ‘velocity’ only appeared in the base position as in ‘wave velocity’, ‘body temperature’ and ‘air resistance’. While the word ‘heat’ only appeared in the head position as in ‘heat capacity’, ‘heat value’. The words ‘speed’, ‘volume’, ‘voltage’, ‘wavelength’, ‘energy’ and ‘pressure’ appeared in both positions in the collocation combinations, thus showing commutability and less collocational restriction than the words ‘temperature’, ‘resistance’, ‘heat’ and ‘velocity’. Restrictions in the position of these words should be noted as these are the lexico-grammatical patterns specific to words used in the English for Science.

Two semantic prosodies found were associated with the semantic sets, type and degree. Semantic prosodies of type and degree were found with the words ‘temperature’ (body temperature, light temperature), ‘energy’ (nuclear energy, high energy), ‘resistance’ (internal resistance, high resistance), ‘speed’ (wave speed, high speed), ‘voltage’ (current voltage, high voltage) and ‘pressure’ (blood pressure, low pressure). The semantic prosody of type was evident with the words ‘volume’ (solution volume) and ‘heat’ (latent heat) while the semantic prosody of degree was evident with the words ‘momentum’ (total momentum, large momentum) and ‘velocity’ (average velocity, uniform velocity). What is noted here is that description of type and degree of certain nouns in scientific texts is preceded by either a noun or an adjective as the collocator, with the keyword in the base position.

Only four of the words registered frequent VN collocations – temperature, heat, energy, pressure. However, similar to the previous semantic group findings, verb phrases formed a small part of the collocations (6%). There were more lexicalized verbs or verbs which carry specific meanings in this group of words, such as ‘produced’, ‘released’, ‘drops’, ‘transferred’, ‘stored’, ‘needed’, ‘dissipated’, ‘consumed’, ‘carried’, ‘supplied’, ‘lose’, ‘changes’, ‘construct’ and ‘conserve’. Most of the lexicalized verbs, however, appear with the word ‘energy’, thus highlighting the need to expose students to the variety of verbs associated with particular scientific words and terms. There were four ‘verb families’ used in combination with the words ‘heat’ and ‘energy’. They were the verb ‘given’ (heat given-53 instances, energy given-7 instances), ‘produce’ (heat produced-10 instances, energy produced-10 instances), ‘released’ (energy released-39 instances, releases energy-12 instances, release energy -6 instances, heat released-12 instances) and ‘absorb’ (absorb heat-8 instances, energy absorbed -6 instances, absorbs energy -5 instances).

3.3 Semantic Category: Related to Forms of Life

Even though the NN and AN combinations were predominant, there were more VN combinations in this set of nouns. Notable was the high percentage of the VN combinations in the words ‘bacteria’, ‘blood’, ‘food’ and ‘organisms’. Even though syntactically they are considered the verb+ing and the ‘verb+ed’ forms, grammatically they perform the role of adjectives, pre-modifying the subject noun such as in the combinations ‘living organisms’, ‘luteinising hormone’, ‘deoxygenated blood’, ‘processed food’, ‘fertilised ovum’, ‘fixing bacteria’ and ‘flowering plants’. In fact the collocation ‘living organisms’ (113 instances) was the most frequent collocation combination in this group.

The words ‘living’ and ‘fixing’ take on extended meanings in these phrases and thus create a more technical level of vocabulary. These are the type of words and word combinations that seem to confuse students as individual word meanings do not apply. Thus, there is much pedagogical relevance attached to the teaching and learning of collocations, especially in the learning of Science subjects.

Three words, ‘cell’, ‘plasma’ and ‘body’, appeared in the head and base positions of the NN combinations showing that these words had less collocational restrictions than the collocations of the other words. The AN syntactic combinations of the keyword ‘human’ displayed a distinct tendency to be formed into 3-word noun phrases, for example, ‘human nervous system’, ‘human respiratory system’, ‘human endocrine system’ and ‘human somatic system’, showing a semantic prosody associated with the type of human systems.

The larger prosodic group found among this semantic set associated with type, as in ‘nucleus cell’, ‘plant cell’, ‘voltaic cell’, ‘human body’, ‘internal body’, ‘pea plants’’ and ‘aquatic plants’. Similar to the findings in the previous sections, collocations referring to the prosodic groups of type and degree normally consist of the NN or AN combinations.

3.4 Semantic Category: Elements on Earth

The syntactic characteristics found here were similar to those discussed in the previous sections. Five nouns: gas, water, air, oxygen and hydrogen, formed less restrictive collocations as the nouns appeared in both the base and head positions of the collocations. However, ‘carbon’, ‘nitrogen’ and ‘chlorine’ appeared only in the head position of the collocations. The collocation ‘carbon atoms’ (162 instances) was the most frequent combination in this group. The modifiers in the AN combinations comprised of both the verb+’ing’ and the verb+’ed’ forms, such as in ‘boiling water’, ‘distilled water’ and ‘compressed air’. This type of formation associated with the semantic prosody of type.

An important finding is that all the words related to gases such as ‘oxygen’, ‘hydrogen’, ‘chlorine’ and ‘nitrogen’ act as the collocator or modifier in the combinations, for example, ‘hydrogen ions’, ‘hydrogen atoms’ ‘oxygen gas’ and ‘chlorine water’. This grammatical pattern could be generalized to other gaseous elements in scientific texts.

Once again, the VN combination formed a small percentage of the collocations, about 6%. There were many lexicalized verbs, such as ‘liberated’, ‘collected’, ‘released’, ‘transports’ and ‘diffuses’. A notable finding is that the verb ‘produce’ and its derivatives was used frequently with the gaseous nouns such as in ‘gas produced’, ‘produce hydrogen’ and produces carbon’. As most of these verbs are different from the verbs used in the previous semantic categories, it can be assumed that these collocations or combinations are fairly restrictive and to some extent, genre specific. However, there are some common verbs used across the semantic categories.

3.5 Semantic Category: Related to Light/Signals/Electricity

All words in this category except for the word ‘electricity’ registered more NN and AN combinations than other types of syntactic combinations. There seem to be more restrictions in the NN collocations of this semantic group, as only the nouns ‘waves’, ‘transformer’ ‘transmission’ and ‘current’ appear in both the head and base positions. The words ‘cathode’, ‘light’ and ‘electricity’ only appear in the head position, as in ‘electricity consumption’, ‘light energy’ and ‘cathode rays’, while the words ‘rays’, ‘anode’, ‘circuit’, ‘coil’, ‘conductor’ and ‘impulses’ appear only in the base position as in ‘gamma rays’ and ‘electric circuit.

Notable was the use of the word ‘light’ as both a noun (red light, light waves) and as an adjective (light temperature, light reaction). Words which are used in different word class forms and appear in collocates in both the head and base positions would be a lexical problem for L2 learners, as they have to be familiar with the variety of ways the word can be used in different contexts.

In comparison to the previous semantic group sets, this group has more instances of the VN combinations, ranging from 3-38%.. Derivatives of the three verbs, ‘produce’, ‘pass’ and ‘generate’, were used frequently in this group. As observed previously, the verb ‘produce’ seems to be used frequently with scientific nouns, thus it could be assumed that ‘produce’ collocates freely with a variety of scientific nouns.

3.6 Semantic Category: Things/Objects/Substances

This group of words could not be given an accurate semantic group. However, as they are not living things and are used to do something, the researcher categorized them under a general semantic category of object/things/substances.

34 out of 54 words in this category had more NN and AN collocations than other types of combinations. The collocations with the keywords in the base position was found to refer to the prosodic group of type as in ‘pure metal’, ‘lactic acid, ‘carbon atom’, ‘biological enzymes’, ‘palm oil’ and ‘fossil fuel’. Notable were two NN combinations - ‘palm oil’ (229 instances) and ‘oil palm’ (105 instances), which had the keyword ‘palm’ in different positions.

23 words registered VN collocations. The verbs linked to the nouns were mainly delexicalised verbs- ‘containing’, ‘needed’, ‘made’, ‘exist’, and ‘using’, which are common verbs used with the nouns in this corpus.

3.7 Semantic Category: General Terms

This group of 34 words consists of concepts and terms which are used across all the Science subjects in this corpus.

The modifiers linked to the nouns in the AN combinations consisted of the verb+'ed' and the verb+'ing' forms such as in the collocations 'dissolved substances', 'colliding particles', 'charged particles', 'pulling force' and 'lifting force'. These are all restricted collocations as the specific adjectives above only link with specific nouns and did not combine with a variety of nouns. The modifiers 'lifting' and 'pulling' formed noun units with extended meanings in the collocations 'Pulling force' and 'lifting force'. The verbs in the VN combinations were limited with most of them being delexicalised verbs, such as ' used', 'found', 'acts', 'consists', 'shows', 'works' and 'produced'.

4. Discussion

4.1 NN and AN Combinations

The collocations in this sample of words, if based on the criterion of commutability, could be assumed to be free collocation combinations. This can be explained through the example of the collocations of the noun 'cell' which in this corpus is seen to collocate with a variety of other nouns both in the head and base positions such as in Figure 1 below.

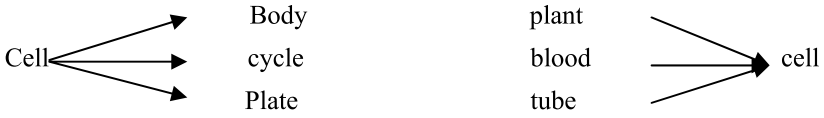

Figure 1. Collocation patterns of the word 'cell'

In the above example, the combination of cell+body and blood+cell, could be assumed to be free combinations, as 'body' also collocates with the word 'human' to form the collocation – 'human body', and 'blood' also collocates with the word 'group' to form the collocation – 'blood group', thus replacing the word 'cell'. Therefore, none of the collocations above seem to be restrictive as to the position or use of the word; many words seem to fit in any one of those positions. This commutability of the elements fulfills the criterion for the collocation to be considered a free combination. However, each element in the collocation does not carry a literal meaning. The combination of cell+body, cell+cycle, blood+cell and tube+cell for example, conveys different meanings, some literal and some non-literal. Thus, most of the collocations found in this corpus fall into the category between free and restricted combinations (Howarth, 1996). The detailed analysis of the collocations have also uncovered clusters of overlapping collocations which show a tendency towards the free end of the restricted category of combinations (Table 2).

Table 2. Collocation clusters

Word Combinations		**Syntactic Combinations**
Head position	**Base position**	
Heat, water	Loss, energy	Noun+noun
Heat, light	Energy, source	Noun+noun
Water, light	Energy, waves	Noun+noun
Water, energy	Loss, content, flow	Noun+noun
Water, air	Molecules, flow	Noun+noun
Water, energy, blood	Loss, flow	Noun+noun
Chemical, new	Cell, substance	Adjective+noun
Excess, less	Heat, water	Adjective+noun
Organic, inorganic	Substance, compounds	Adjective+noun
Constant, final, average	Temperature, velocity	Adjective+noun

Average, critical	Temperature, mass	Adjective+noun
High, higher, low, external	Temperature, pressure	Adjective+noun
Harmful, certain	Bacteria, microorganisms	Adjective+noun
Normal, slow	Growth, reaction	Adjective+noun
Genetic, chemical	Factors, composition	Adjective+noun
Chemical, ionic	Equations, bonds	Adjective+noun

The collocations, in Table 2, show that the elements or words in the head positions freely combine with a number of words in the base positions, as depicted in Figure 2. This type of flexible combinations can help in prediction of other collocational combinations.

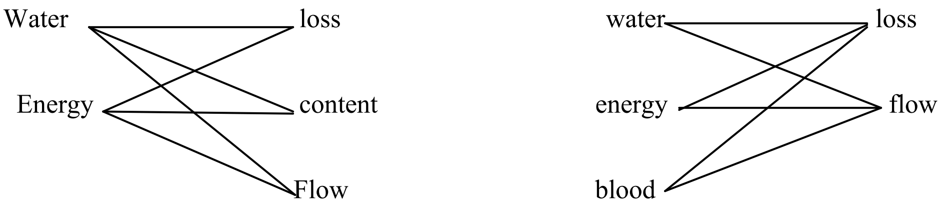

Figure 2. Free combinations

However, in the flexibility there are constraints. The above flexibility in combinations cannot be applied to every noun of the same lexical set such as in the collocations of the words ‘air’, ‘flow’, ‘heat’ and ‘loss’ (Figure 3).

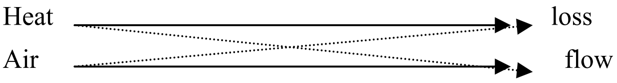

Figure 3. Collocational restrictions

The collocations ‘heat loss’ and ‘air flow’ are possible but not ‘air loss’ and ‘heat flow’. The prediction in this case is arbitrarily blocked by usage (Howarth, 1998). These types of collocational clusters are the type of phenomenon which might confuse learners.

4.2 VN Combinations

The more common type of verbs which appeared in the VN collocations were lexicalized verbs or verbs which carry contextual meaning such as ‘flows’, ‘produce’ and ‘enters’. Some verbs such as ‘produce’, ‘released’ and ‘cause’ were seen to be frequently used across the corpus.

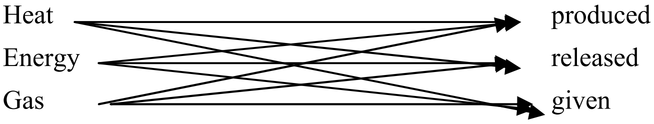

Figure 4. Overlapping VN clusters

Even though, there were various overlapping collocational clusters of the NN and AN combinations, there seem to be only one set of overlapping collocational cluster of the VN combination, as presented in Figure 4. This finding shows that the VN collocations in the Science corpus lean more towards the restricted end of the collocation categories (Cowie, 1994).

5. Conclusion

This study shows the importance of learning language chunks or collocations. The predictable flexible combinations of the noun keyword collocations could help students to predict other types of collocational combinations. The predictability of collocations is also helped by identifying the semantic prosodies of the combinations as the combinations with the same grammatical pattern seem to share aspects of meaning. These patterns can be taught explicitly to students so as to prepare them for possible collocations, once the prosodic group has been identified.

However, these patterns should not be over-represented. Students should only be informed of the typical patterns but reminded of the diversity and changeability of collocation patterns in Science as the flexibility of some combinations is arbitrarily blocked by usage. This proves that even though meaning and pattern are associated in the language used in Science, the extent or limitation of that association cannot always be specified. This arbitrary lexical patterning or phraseology is an area which would pose enormous problems to students and is an area that needs to be considered when developing materials for any EAP course. This however, cannot be done without reference to a relevant corpus such as the one created in this work.

References

Atkinson, D. (1999). Language and Science. *Language and Science Annual Review of Applied Linguistics*, *19,* 193-214. Retrieved from http://www.uefap.com/writing/research/langsci.htm. http://dx.doi.org/10.1017/S026719059919010X

Barber, C. L. (1962). Some measurable characteristics of modern scientific prose. In Contributions to English Syntax and Philology. Reprinted in J. M. Swales (Ed). (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 from http://acl.ldc.upenn.edu/W/WOO/WOO-0902.pdf

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, *2,* 223-235. http://dx.doi.org/10.1093/applin/2.3.223

Cowie, A. P. (1994). Phraseology. In R.E. Asher (Ed.) *The encyclopedia of language and linguistics* (pp. 3168-3171). Oxford: Pergamon Press.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for Academic purposes. *English for Specific Purposes, 28*(30), 157-169. http://dx.doi.org/10.1016/j.esp.2009.02.002

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), Researching pedagogic tasks: Second language learning, teaching and testing (pp. 75–93). Harlow, UK: Longman.

Gledhill, C. (1996). Collocations and the rhetoric of scientific ideas. Corpus linguistics as a methodology for genre analysis. Retrieved from http://gandalf.aksis.uib.no/allc/gledhill.pdf.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115-135. http://dx.doi.org/10.1016/S0889-4906(98)00015-5

Herbert, A. J. (1965). The structure of technical English. In J. Swales (Ed), *Episodes in ESP*, (pp. 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.

Hill, J. (2000). Collocation: Practical classroom issues. In M. Lewis (Ed.) *Teaching collocations*. Hove: Language Teaching Publications.

Hoey, M. (2007). Lexical priming and literacy creativity. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Series editors), *Text, discourse and corpora: Theory and analysis* (pp. 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.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. http://dx.doi.org/10.1093/applin/19.1.24

Hunston, S., & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes, 13,* 239-256. http://dx.doi.org/10.1016/0889-4906(94)90004-3

Johnson, S., & Ensslin, A. (2006). Language in the news: Some reflections on keyword analysis using WordSmith Tools and the BNC. *Leeds Working Papers in Linguistics and Phonetics*, *11*. Retrieved from http://www.leeds.ac.uk/linguistics/WPL/WPL11.html.

Johnson, S., Culpeper, J., & 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. http://dx.doi.org/10.1177/0957926503014001928

Kjellmer, G. (1991). A mint of phrase. In K. Aijmer & B. Altenberg (Eds), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111-127). London: Longman.

Kornwipa, P., Somcheon, H. J., & 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.

Lackstrom, J. E., Selinker, L., & Trimble, L. (1973). Technical rhetorical principles and grammatical choice. *TESOL Quarterly*, *7*, 127-136. http://dx.doi.org/10.2307/3585556

Lewis, M. (1993). *The lexical approach*. Hove, England: LTP.

Lewis, M. (2001). *Implementing the lexical approach*. *Putting theory into practice*. 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. http://dx.doi.org/10.1016/S0889-4906(98)00013-1

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

Moon, R. (1998). *Fixed expressions and idioms in English. A corpus-based approach*. Oxford: OUP.

Mudraya, O. (2006). Engineering English: A lexical frequency instruction model. *English for Specific Purposes*, *25*, 235-256. http://dx.doi.org/10.1016/j.esp.2005.05.002

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

Nattinger, J. R. (1988). Some current trends in vocabulary. In R. Carter & M. McCarthy, (Eds) (1988), *Vocabulary and language teaching*. London/New York: Longman.

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

Nesselhauf, Nadja. (2004). What are collocations? In David Allerton, Nadja Nesselhauf & Paul Skandera (Eds). *Phraseological Units: Basic Concepts and Their Application* (pp. 1-21). Basel: Schwabe.

Nesselhauf, Nadja. (2005). *Collocations in a learner corpus*. Philadelphia: John Benjamins.

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

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

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Richards, J. C. & Schmitt, R.W. (Eds) *Language and Communication*, (pp. 191-225). London: Longman.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. In N. Schmitt (Ed), *Formulaic Sequences* (pp.1-22). Amsterdam: John Benjamins.

Scott, M. (2000). Focusing on the text and its key words. In L. Burnard & 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, & 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 & 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.

Sinclair, J. M. (1996). The empty lexicon. *International Journal of Corpus Linguistics*, *1*(1), 99-119. http://dx.doi.org/10.1075/ijcl.1.1.07sin

Skehan, P. (1998). A cognitive approach to language learning. 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 Purposes*, *21*(2), 145-165. http://dx.doi.org/10.1016/S0889-4906(00)00034-X

Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23-55. http://dx.doi.org/10.1075/fol.2.1.03stu

Swales, J. (1998). *Other floors, other voices: A textography of a small university building*. Mahwah, NJ: L. Erlbaum.

Tarone, E., Dwyer, S. Gillette, S., & 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. http://dx.doi.org/10.1016/0272-2380(81)90004-4

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

Trimble, Louis. (1985). *English for Science and technology: A discourse approach*. New York: Cambridge University Press.

Ward, J. (2007). "Collocation and technicality in EAP engineering". *Journal of English for AcademicPurposes, 6,* 18-35. http://dx.doi.org/10.1016/j.jeap.2006.10.001

Ward, J. (2009). "A basic engineering English word list for less proficient foundation engineering undergraduates". *English for Specific Purposes, 28,* 170-182.

Williams, G. C. (1998). "Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles". *International Journal of Corpus Linguistics, 3,* 151-171. http://dx.doi.org/10.1075/ijcl.3.1.07wil

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, *32*(4), 213-231. http://dx.doi.org/10.1017/S0261444800014154

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511519772