11
CORPUS ANALYSIS OF PRIMARY ONE SCIENCE TEXTBOOKS FOR DESIGNING ELT MATERIALS

SARIMAH SHAMSUDIN
ZAIDAH ZAINAL
SALBIAH SELIMAN
YASMIN HANAFI ZAID

INTRODUCTION

To make use of words or lists of words in various forms for various purposes is not new. We have been using lists of vocabulary words for tourists and students from various levels of education in the form of glossaries, lists of jargons, indexes and the like. Such lists are called corpus.

The government's recent policy on the teaching of Science in English calls for a fundamental support from language practitioners and researchers of these fields. Here, we highlight some important issues regarding the use of English as the medium of instruction for the teaching and learning of Science in primary schools. Among others, the language issue related to the lexical, syntactic and semantic patterns of English in Science and Technology (EST) has been 'under-researched'. This, therefore, sets the focus of our study which undertakes to examine the language patterns existing in science authentic texts. Among the many conventional methods that can be adopted, such as functional-notional @ communicative method (Wilkins, 1976), structural @grammar approach (Chomsky, 1965), procedural approach (Prabhu, 1987) or instrospective and retrospective methods (Pressley and Afflerbach, 1995) which often

times are limited and unsystematic, we propose to employ the method which involves the making of corpus of this subject area using lexical approach (Lewis, 1994). The lexical approach (Lewis, 1993; Willis, 1990, Willis & Willis, 1988, 1989) is chosen for a number of reasons: 1) it emphasises on the importance of co-text (i.e. language is not de-contextualised), 2) it provides a range of awareness-raising activities that direct the learner's attention to chunks text composed, 3) it focuses on different forms of lexical item. The corpus produced can then be used by other researchers in this area for teaching and learning purposes.

In this paper, we will discuss the preliminary stage of an on-going research which aims to design teaching and learning materials through an analysis of a corpus of texts taken from Science textbooks for Primary One students in Malaysia. The topic of our research is 'EST Teaching and Learning Materials via WWW Based on Corpus Analysis of Mathematics, Science and English Text Books in Malaysian Primary Schools'. This paper, however, only focuses on the use of the frequency list and corpus of Science texts to develop teaching and learning materials for English language learners of Primary One students.

CORPUS AND CONCORDANCES

What is a corpus? A corpus is defined as 'a computer-based text or collection of texts' (Barlow, 2003) in machine-readable form. It is 'any text-only file, or set of text-only files that can be loaded into the program (concordancing program) for analysis' (Barlow, 2003: 19). In linguistic settings, Sinclair et al. (1987) defined corpus as 'a collection of texts, of the written or spoken word, which is stored and processed on computer for the purpose of linguistic research'.

Today, there has been a growing interest and need in corpus making and use. These interests, among others, are based on the following reasons (Svartvik, 1992: 8-9):

1. language practitioners require large amounts of easily accessible and authentic data. Language complexities may not be sufficiently captured by introspection and elicitation alone. Corpus, in comparison to random introspective observation, is deemed necessary to objectively examine data.
2. corpus is necessary in describing the different language uses and in establishing between frequency of occurrence of linguistic items.
3. corpus is a general source of information for indexing key words and concepts.
4. some subject areas, e.g. science and technology, adhere to specific language styles, functions and formats. These can be captured through corpus analysis.

Corpus is, therefore, a collection of naturally-occurring language text, chosen to characterize a language variety or discourse. As the lists of words are often extensive and bulky, analysis of corpus through the use of software such as Monoconc is sought. Monoconc enables language practitioners to instantly work out the frequency list of words and phrases in the corpus and patterns of their occurrence. Such patterns of occurrence can be in the form of concordance.

What is ***Concordance***? Sinclair (1991:32) defines a concordance as "a collection of the occurrences of a word-form, each in its own textual environment". It is basically an index to the words in a text. Concordancing is the technique of locating the occurrences of a specific word that has been selected in a particular corpus and consequently listing its context very quickly and reliably. In other words, concordance gives access to important language patterns in texts and finds key words in context, i.e. words that occur on either side of a word or phrase selected for study.

It gives access to important language patterns in texts and finds key words in context, i.e. words that occur on either side of a word or phrase selected for study. An example of the concordance

of the word *pupils* from a corpus of science texts for Primary One students in Malaysia is shown in Figure 1 below.

Concordance is also a Co-text. The Co-text of a selected word or phrase consists of the other words on either side of it. For instance in Figure 1 below, the Co-text of the word *pupils* are *get-to, let-know, ask-to/which* etc.

Figure 1: Co-text of the word *pupils* [Concordance of *pupils*]

CONCORDANCE

Concordancing software or computer concordancers such as MonoConc Pro 2.2 (MP 2.2) can be used to rapidly search for patterns in a corpus using its search query. It can be used to analyse lexical,

grammatical and textual structures of a corpus.

Concordance has several advantages. For instance, by using the software Monoconc Pro 2.2 (MP2.2), it is possible to search for

1. frequency lists of word occurrences
2. rare instances of words or strings of words
3. strings of words in the context of other strings
4. particular patterns of words and sorts them to focus on similar occurrences to reveal their properties

Corpus analysis studies have been numerous (see Tribble and Jones, 1997; Murison-Bowie, 1993; Willis and Willis, 1988, 1989, among others). For instance, Willis and Willis (1988, 1989) attempt to design course books for English language learners through authentic evidence found in the COBUILD corpus. Most useful words and patterns are identified and presented in the course books to English language learners in order to give them a good start with instances of real and most frequent patterns of the target language.

An ESP web-based courseware called UNITEKMA ECE courseware for Civil Engineering students at the tertiary level of education has also been designed based on computational linguistic analysis of a corpus on Civil Engineering materials (Sarimah Shamsudin, 1997; Sarimah Shamsudin et al. 2002) Research conducted on the use of the web-based ECE Courseware indicated that the courseware is able to provide students with authentic and real examples of the language in the context of Civil Engineering materials. The glossary available in the courseware helps the students gain new vocabulary and definitions of terms and concepts in civil engineering area.

Currently, there are many different collections of corpora of English, especially general written and spoken corpus. A few examples are the Collins Birmingham University International Language Database (COBUILD) of English, Lancaster-Oslo-Bergen (LOB) Corpus of British English and the Brown Corpus of American English.

METHODOLOGY OF THE ONGOING RESEARCH

Let us begin to describe the on-going research which aims to design teaching and learning materials through an analysis of a corpus of Science textbooks for Primary One students in Malaysian schools. The main study intends to include texts from the areas of Science, Mathematics and English primary one textbooks.

First, text books and workbooks in the respective subjects used in Primary Schools in Malaysia were gathered and copyright permission was sought out. Dewan Bahasa dan Pustaka (DBP) was very prompt in giving the copyright of the titles they published. The other publisher has not given the permission to date although permission was sought at the same time with DBP.

For the purpose of this presentation, only Science text books were analysed. The procedures taken to get the data for this paper were similar to those that we took for the whole research and they were as follows:

1. each page of the science textbook was scanned using C-pen 10 and saved as textfiles
2. All the textfiles for science texts were merged
3. the merged textfile or corpus was analysed using Monoconc to get the results that are useful for our purpose

A frequency list was produced using MonoConc (See Figure 2). Below is the first page of the actual frequency list produced. A complete frequency list is found in Appendix A.

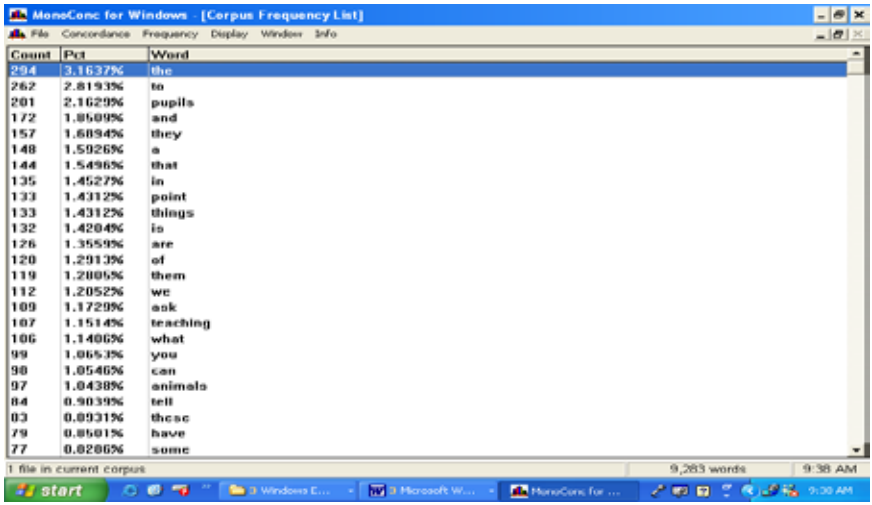

Figure 2: Frequency list

The list produced mixed results and the words that were deemed frequent, can then be classified into grammatical or structural words and content words, as shown in Table 2 below.

Grammar Words		Content Words	
Frequency	Words	Frequency	Words
294	The	201	pupils
262	To	133	point
172	And	133	things
157	They	109	ask
148	A	107	teaching
144	That	97	animals
135	In	55	see
132	Is	52	sound
126	Are	48	make

120	Of	46	water
119	Them	44	eat
112	We	44	plants
106	What	44	food

Table 2: Frequency of Grammar and Content words

From frequency lists (Appendix A, Table 2) and their concordance, teachers can develop learning and teaching materials for English language learners.

DEVELOPING TEACHING AND LEARNING MATERIALS FOR ENGLISH LANGUAGE LEARNERS

This paper further discusses how to use the frequency list and corpus of Science texts to develop teaching and learning materials for English language learners of Primary One students.

For instance in Appendix A, the content word ***things*** is one of the common words in the corpus. It can be considered as an important concept for Primary One learners of Science as its occurrence in the corpus consists of 9 242 words is 133…..

In addition, we may design exercises based on the **collocation** of the words found in the frequency list such as things (frequency: 133 - Figure 4) and *animals* (frequency: 97 – Figure 3). To **collocate** is to find the word which occurs in close proximity to the word under investigation. Thus a **collocation** is the occurrence of two or more words within a short space of each other in a text. Usual measures of proximity are a maximum of four words intervening.

Figure 3: The collocation of the word *animals*

The frequency list (appendix A) determines which word we need to collocate. Our decision should be based on the necessity of the word. The more the frequency of occurrence, the more important the word will be. After the word is decided on, it is run through the MonoConc to work out the collocation, ie the context the word actually appears in the science texts analyzed. For instance the word *things* is found to be frequent. Figure 3 below shows the collocation of the word *things*.

Figure 4: The collocation of the word *things*

From the following figures we can see that the words surrounding the target word *things* include *the*, *then* and the like. Nevertheless, the most important information that we should know here is the co-text of the word in question.

Figure 5: The collocation of *big things*

The words *big* and *small* seem to be very salient with the word *things* (See figure 9).

Figure 6: The collocation of *small things*

After we found out that the words *small* and *big* were very salient with the words *things*, we were able to use the same method to work out the collocation of these words. The result of the collocation for these two words with the word *thing* can be expressed in Figure 7 and 8. From Figure 7 we learnt that the next salient word to *small* is the word *float* and from Figure 8 we learn that the next salient word for big things is *sink*.

Figure 7: The collocation of *float*

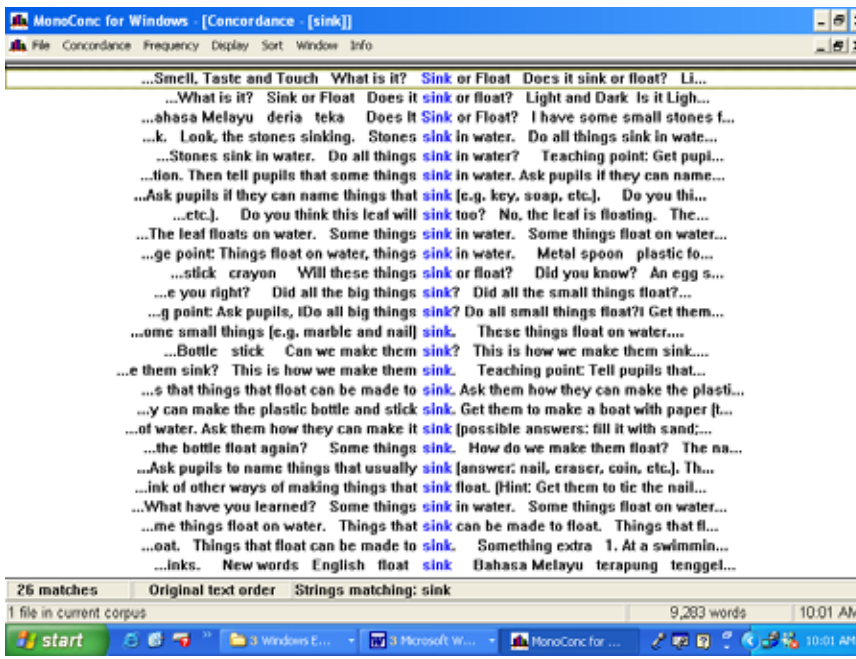

Figure 8: The collocation of *sink*

The same method can yet be used to find out what were referred to by the word *things*. Figure 9 lists some examples of words which were meant by *things*.

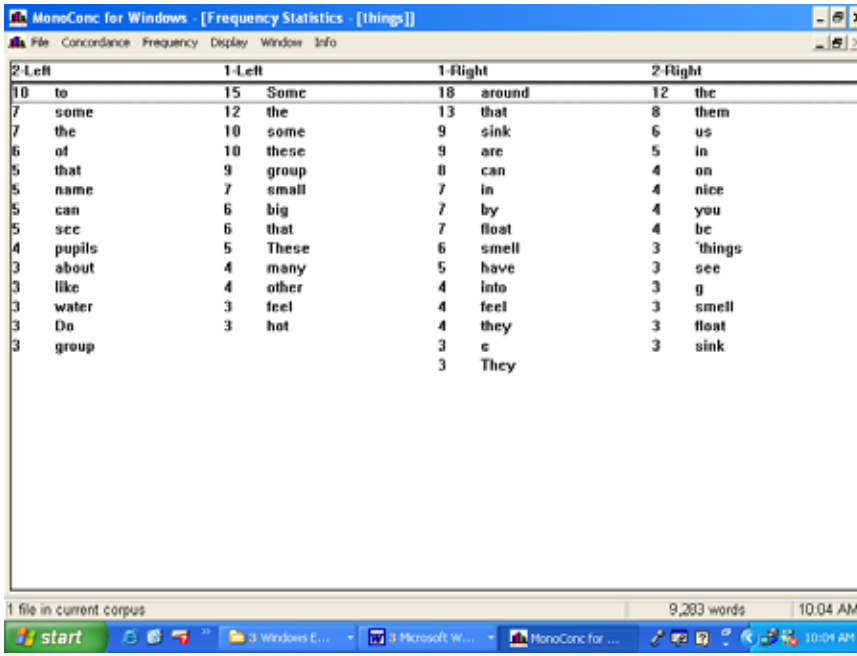

Figure 9: The frequency of those we can call *things*

With all these findings we can then teach the concept of things and other salient words that collocate with it such as *big, small, float* and *sink*.

Below are some sample exercises that teachers can try:

Sample Exercise 1:

Prepare either cut out pictures (low proficiency) or words (higher proficiency) that can stick on a board (e.g. metal board) by means of magnetic strip. These cut-out pictures or words consist of those we can call *things*. Prepare the word *things* that will be put in the middle of the board. The cut-out pictures or words are placed on a table so that students can see them. Students are asked to place the correct pictures or words. A completed student activity will look like Figures 10 and 11.

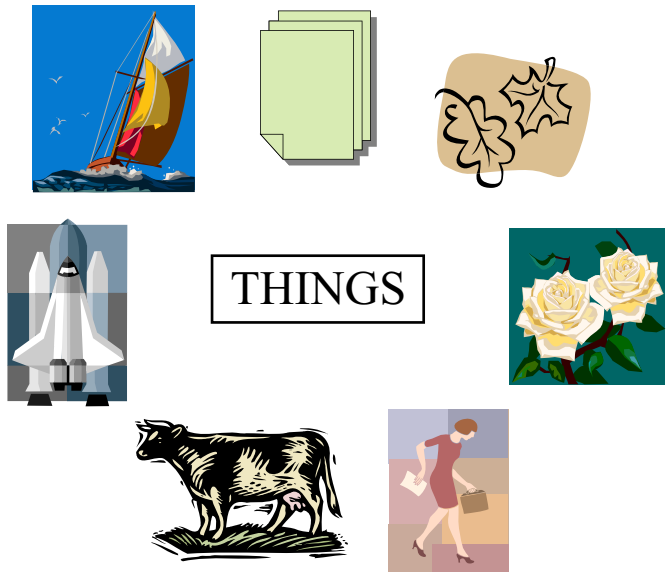

Figure 10: Board 1a: pictures

spaceship	flower	leaf
boat	THINGS	animal
paper		people

Figure 11: Board 1b: words

Sample Exercise 2:

Words that can qualify as *things* included in the previous exercise, such as *big, small, sink, float* and the like, can now be added. With this new addition, students will need to identify these words in relation to the earlier ones. Figure 12 illustrates this by using the words sink and float. Teachers can also use big and small in the same manner.

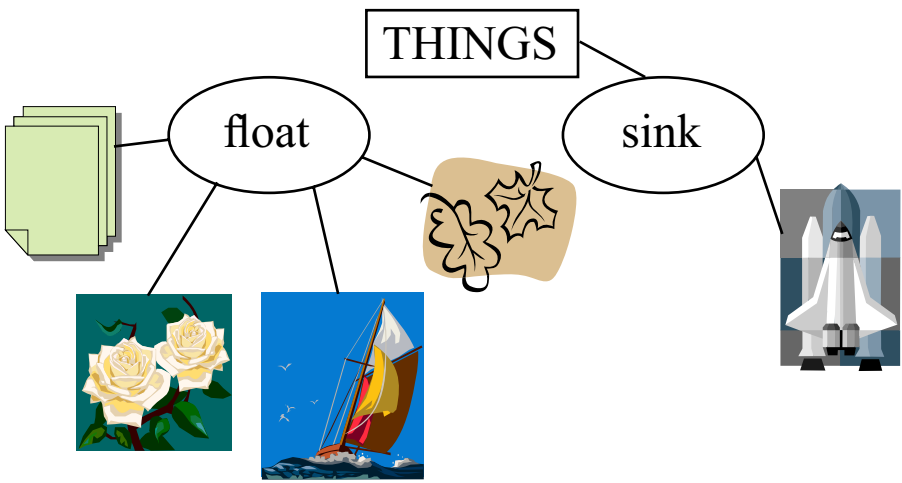

Figure 12: Board 1a: sink and float - pictures

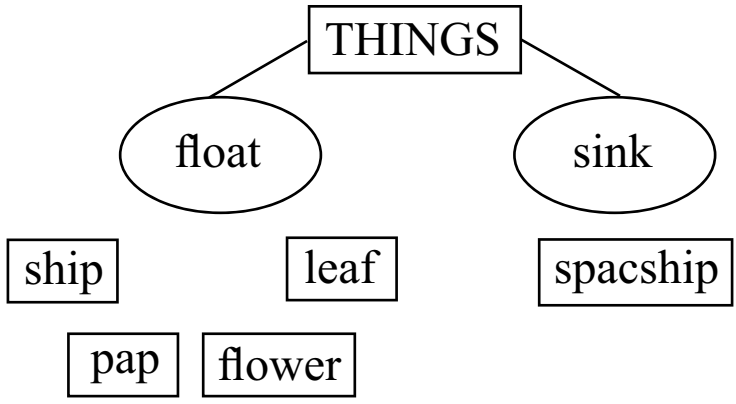

Figure 13: Board 1a: sink and float - words

The organisation of the game can be done in two ways. The first way is to use magnetic strip stuck at the back of all cut-out words or pictures. (See figures 10, 11 and 13). The students can move these words or cutout pictures as they need to. The second way is to use string. A piece of string starts from the word *things* (See figure 12). The string can be used by students to join the words for various exercises.

CONCLUSION

From this preliminary research, we can conclude that corpus analysis of texts, even at the primary school level, would be able to provide us with new opportunities in designing teaching and learning materials to be used in schools. Corpus gives teachers and materials writers, new perspectives in their work. With the help of computer and computer

programmes, the job of analyzing words and phrases becomes easier, more accurate and more interesting. Those lists of words can be regarded as important, as they are obtained through the analysis of corpus, rather than doing it by intuition, as done by most text or workbook writers. Teachers should take the opportunity to try such means. Our research tries to produce materials which will enable students learning the English language to benefit from the policy of teaching science and mathematics in English.

In our research, we also would like to find out if students who undergo corpus based materials have better understanding of generic concepts in their chosen fields.

REFERENCES

Aston, G., (2001) *Learning with Corpora* Houston: Athelstan.

Barlow, M., (1997) *MonoConc for Windows* Houston: Athelstan.

Barlow, M. (1997). A Guide to Monoconc. http://www.ruf.rice.edu/~barlow/mc.html

Barlow, M., (2003) *Concordancing and Corpus Analysis Using MP2.2* Houston: Athelstan.

Chomsky, n. (1965) *Aspects of the Theory of Syntax*. Cambridge, Mass: Cambridge University Press.

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

Murison-Bowie, S. (1993) . *MicroConcord Manual: An Introduction to the Practices and Principles of Concordancing in Language Teaching*. Oxford: Oxford University Press.

Prabhu, N. S. (1987). *Second Language Pedagogy*. Oxford: Oxford University Press.

Pressley, M. and Afflerbach, P. (1995). *Verbal Protocols of Reading*.Lawrence Erlbaum Associates, UK.

Salbiah Seliman, Zaidah Zainal, Sarimah Shamsudin, Yasmin Hanafi Zaid. (in progress). *EST Teaching and Learning*

Materials via WWW Based on Corpus Analysis of Mathematics, Science ***and English Text Books Used in Malaysian Primary Schools***. IRPA Project 74234.

Sarimah Shamsudin (1997). *Introducing Self-Access ESP CALL Material Based on Corpus Analysis Via the World Wide Web for the English for Civil Engineering Programme in University Technology Malaysia*. Unpublished Masters Dissertation, Aston University, Birmingham, UK.

Sarimah Shamsudin et al. (2002). *Courseware Development on Civil Engineering Construction Materials for Self-Access Language Learning via the WWW:A Computational Linguistic Analysis Approach*. A report for the Research Management Centre, Universiti Teknologi Malaysia.

Sinclair, J. et al. (1987). *Collins Cobuild English Language Dictionary*. London: William Collins Sons & Co. Ltd.

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

Svartvik, J. (1992). Corpus linguistic comes of age. In Jan Svartvik (ed.) *Trends in Linguistics Studies and Monographs 65: Directions in Corpus Linguistics*. Mouton De Gruyter.

Tribble, C and Jones, G., (1989) *Concordances in the Classroom: Resource Book for Teachers* Essex: Longman Group UK Limited.

Wilkins, D. (1976). *Notional Syllabuses*. Oxford: Oxford University Press.

Willis, D. (1990). *The Lexical Syllabus*. London: Harper Collins Publishers.

Willis, J. & Willis D. (1988). ***Collins Cobuild English Course: Students' Book 1***. London: William Collins Sons & Co. Ltd.

Willis, J. & Willis D. (1989). *Collins Cobuild English Course: Students' Book 1*. London: William Collins Sons & Co. Ltd.