

## **CHAPTER II**

### **LITERATURE REVIEW**

#### **2.1 Introduction**

In this chapter, the literature related to the study is reviewed in order to obtain a theoretical framework for implementing the concordance-based method in comparison with the conventional teaching method on vocabulary learning. The review is divided into four main areas: academic vocabulary, the nature of vocabulary acquisition, vocabulary instruction and the concordance-based method. Firstly, the academic vocabulary is identified in terms of vocabulary types, lexical thresholds for academic reading, and assessment of students' vocabulary size. Principles derived from reviewing these matters are used in the present study as criteria for selecting target words systematically as well as designing the tests for assessing the vocabulary in focus. Secondly, two types of knowledge being focused in the study i.e. definitional knowledge and transferable knowledge are described and the nature of vocabulary acquisition and retention is examined in order to derive insights how to enhance students' vocabulary acquisition and retention. .

Thirdly, the background of vocabulary instruction is presented to highlight the importance of vocabulary instruction at present. After that, the main current approaches to vocabulary instruction are reviewed in order to indicate the strengths of each method and their compatibility to the design of the concordance-based lessons. Fourthly, the concordance-based method is introduced in terms of its background, description and learning approach of DDL before its roles in ELT are discussed in terms of its compatibility to various approaches to vocabulary instructions and previous applications. Finally, the status of the present study in this area of research is established.

#### **2.2 Academic Vocabulary**

In general, lexical knowledge is accepted as a necessary factor for successful reading comprehension. The more the readers know about words in a particular

passage, the better they can comprehend that passage. To increase students' lexical knowledge for academic reading as in the present study, however, the exact words must be identified in order to ensure that they are worth being studied. Accordingly, considerations must be taken towards vocabulary types and lexical thresholds for academic reading in order to obtain the principles for selecting words most suitable for the students' needs. Apart from word selection, the measure of such word knowledge is also important for accurate assessment. As a result, word selection and instrument design for assessing vocabulary size in the study are based on previous works in the following areas: types of vocabulary, lexical thresholds for academic reading, and assessment of vocabulary size.

### **2.2.1 Types of vocabulary**

It is quite obvious that different sets of vocabulary are needed for different types of language use. In reading, for example, some vocabulary is found more frequently in textbooks than in newspapers or advertisements whereas some is used more often in engineering textbooks than in business or medical ones. In teaching academic reading, it is, therefore, advantageous to pay attention to words frequently occurring in students' academic texts and closely relevant to their needs. The lessons will be more meaningful if the vocabulary that students are likely to encounter in their future study and career is taught. Several studies have investigated the vocabulary needed for academic reading. One such attempt has been conducted in the form of a frequency-based method. In these studies (Nation, 2001), words found in a very wide range of text types were counted according to the frequency of their occurrences and then identified and classified for pedagogical use. Based on these studies, vocabulary is divided into four main types i.e. high frequency words, academic words, technical words and low frequency words (Nation, 2001; Coxhead and Nation, 2001; Dudley-Evans and St. John, 1998; and Nation and Waring, 1997).

#### **2.2.1.1 High frequency words**

A group of high frequency words is regarded as an essential basis for all language use. This group consists of around 2,000 word families, which covers a very large proportion of the running words in spoken and written texts and occurs in all kinds of language use. Some examples of high frequency words are

'answer, work, high, idea, left, metal, often, page, mile, strong, usually' etc. All function words such as prepositions, conjunctions and auxiliary verbs are also included in this group. Many lists of the most frequently occurring words in English are established for convenient use in pedagogy, but the most commonly cited list is the classic collection of Michael West (1953), *The General Service List of English Words (GSL)*. Nation and Waring (1997, p.15) mention, "Although the GSL is in need of replacement because of its age, errors it contains, and its written focus, it is still the best available list, given a range of information it contains about the relative frequency of the meanings of the words." Several studies indicate that the GSL can provide coverage of around 80% of the running words in most academic texts (Coxhead and Nation, 2001; Cobb and Horst 2001; and Nation and Waring, 1997). This means that when readers have control of the 2,000 words of general usefulness in English, they should be able to understand 80% of texts i.e. two words unknown per printed line or one unknown word in every five words.

#### **2.2.1.2 Academic words**

Academic vocabulary is a group of words occurring frequently over a wide range of academic texts across disciplines i.e. not restricted in any specific domains. Generally, these words are not so common in non-academic texts, and they are not technical terms in any particular domain either. Instead, such words are usually found over a range of academic texts or formal papers such as secondary-school and university textbooks, specialized journals, reports, manuals, or newspapers etc. Academic vocabulary is sometimes referred to as '*semi-technical vocabulary, sub-technical vocabulary, or specialized non-technical lexis*', (Coxhead and Nation, 2001). Some examples of academic vocabulary are '*assume, cite, capable, approach, aspect, crucial, element, feature, integrate, justify, manipulate, vision, publish, accurate*' etc. Two outstanding lists of academic vocabulary are *The University Word List (UWL)* (Xue and Nation, 1984; and Nation, 1990), and *The Academic Word List (AWL)* (Coxhead, 1998).

The UWL is a compilation of four separate studies, according to Coxhead and Nation (2001) and Nation and Waring (1997). The UWL consists of around 800 word families, not included in the first 2,000 words of the GSL. It is found from an analysis of text corpora in a variety of disciplines that the GSL together

with the UWL reliably account for 90% of tokens. In other words, the inclusion of the UWL knowledge to the GSL knowledge can increase 10% of the text coverage i.e. from 80% to 90%. More recently, the AWL has been similarly developed to see which words in the list are truly academic words and which are general service words, not in West's GSL. Coxhead and Nation (2001, p.254) mention that the list of 570 AWL words provides slightly better coverage of academic texts than the UWL even though it contains fewer words. As a result, knowing the GSL together with the UWL or the AWL will also give close to 90% coverage of the running words in most academic texts.

### 2.2.1.3 Technical words

Technical words are very closely related to the topic and subject area of the texts. They are reasonably common in one topic area but not so common in the others. For example, the words '*phoneme, morpheme, lemma, hapax legomena*' are restricted in Applied Linguistics whereas the words '*anode, impedance, dielectric, galvanometer*' are exclusively used in Electronics. Therefore, these types of words differ from one subject area to another. In some cases, the same words are used in various fields, but with different meanings. For instance, the words '*operation*' in the Medical field, '*mouse*' in Computer Science, '*strength*' in Physics, and '*overhead*' in Business mean something different from those in general or in other fields. Similarly, the words '*range and frequency*' in Linguistics have a completely different meaning from those in Electronics. However, technical words in any particular subject are probably about 1,000 words or less as it can be noticed from the number of headwords in any specialized dictionary. These words typically cover about 5% of the running words in a text, according to Coxhead and Nation (2001). With the text coverage of 5%, knowledge of technical words can enhance that of high frequency and academic words to get closer to the 95% threshold of text coverage.

### 2.2.1.4 Low frequency words

This group consists of words typically occurring in a very narrow range and low frequency. So far, this group is the biggest since they are all words which are not included in the above three groups. These words, for example, include proper nouns, words which almost belong in the high-frequency list, words

rarely found in language use, words found only once or twice in one text and seldom appearing in other texts, etc. The study of Carroll, Davies and Richman (1971, cited in Coxhead and Nation, 2001) found that 40.4% of 5,000,000 running words in a corpus were 86,741 different word types occurring only once or twice.

This classification of words is certainly useful for pedagogical practice because it provides a clear goal for teaching vocabulary. It suggests the type as well as the number of words that need to be learned so as to be able to cope effectively with specific goals in language use. It is quite obvious that the most frequently occurring words should be primarily dealt with. From these four types of vocabulary, it has been agreed that at least the GSL 2,000 high frequency word families must be learned by foreign language learners in order for them to have an essential basis for all language use (Cobb and Horst, 2001; Coxhead and Nation, 2001; and Nation and Waring, 1997). For academic study, however, these 2,000 words are still insufficient to empower readers. If English is to be used for academic study, general academic vocabulary i.e. the UWL or the AWL of about 500-800 word families must be added to the knowledge of general vocabulary. After gaining control of the GSL 2,000 high frequency words, then learners are suggested to focus on general academic vocabulary. Cobb and Horst (2001, p.319) point out that the two groups of high frequency words and academic words “constitute a general English for Academic Purposes (EAP) vocabulary syllabus that takes a learner to the outer edge of reading in a specific domain”. In addition, Coxhead and Nation (2001, p.260) confirm that knowing these two groups of words “will give close to 90% coverage of the running words in most academic texts. When this is supplemented by proper nouns and technical vocabulary, learners will approach the critical 95% coverage threshold needed for reading”. Therefore, it is often suggested that the GSL together with either the UWL or the AWL is the minimum lexical knowledge base for reading in any academic domain. In other words, it has been estimated that ‘a lexical threshold’ to reading comprehension of academic texts should be around 3,000 word families including about 2,000 high frequency words in the GSL and 500-800 academic words in the AWL or the UWL.

On the other hand, knowledge of technical words largely depends on existing or specialized knowledge whereas low frequency words are in a large number of

different word types, but with rare occurrences in a piece of text. Technical vocabulary is sometimes not considered to be the English teacher's job to teach (Coxhead and Nation, 2001), since it seems to be less of a problem to the learners and it can be naturally acquired from their specialized courses. Regarding low frequency words, although this type of vocabulary is in a large group, most of them occur only once or twice in a piece of text. Despite a low coverage of text i.e. 5%, they are too large in number to all be learnt in academic courses. Therefore, it is not practical to spend most of the course time learning words which are occasionally found in the target texts. To prepare students to deal with these two types of words, the students should be taught reading strategies such as guessing word meanings from context clues, analyzing word parts, or using word cards and dictionaries (Nation, 2001; and Coxhead and Nation, 2001). In specialized texts, for example, when technical terms are introduced, they are normally defined and exemplified. Therefore, it is a good idea to teach the students to cope with texts by using the context clues of definitions and examples.

Regarding the present study, it is obvious that the first two word types i.e. high frequency and academic words are very suitable for the students because they are necessary for academic reading. As a result, one criteria of word selection in the study are based on the established wordlists of high frequency words (the GSL) and academic words (the AWL). These wordlists are used as '*reference lists*' for selecting words from the corpus.

### **2.2.2 Lexical thresholds for academic reading**

Pedagogically, a lexical threshold for academic reading is the estimate of the minimal lexical knowledge used as a critical basis for academic reading. It is a useful criterion in setting teaching goals, diagnosing students' weaknesses, and designing syllabuses or lessons. Lexical thresholds are described in the areas of reading comprehension and vocabulary size in order to identify which words can be used as critical basis for academic reading in the present study.

#### **2.2.2.1 Lexical threshold for reading comprehension**

In previous studies of reading comprehension, one important area to investigate is the relationship between vocabulary coverage and reading

comprehension. The main interest is to define the minimal language lexical knowledge to be transferred to reading comprehension. Accordingly, the percentage of known and unknown words was calculated in order to determine the coverage of texts sufficient for comprehending that text. Very interesting findings were revealed by Laufer's studies (cited in Nation, 2001; and Cobb and Horst, 2001) conducted in 1989 and 1992. It was found from her studies that vocabulary coverage correlated consistently with reading comprehension. Students with scores of 95% and above on the vocabulary measure were significantly more successful on the reading measure than those scoring below 95%. Therefore, Laufer determined such percent coverage as minimally acceptable comprehension. In determining vocabulary size for providing 95% coverage of academic texts, her studies consistently showed that the 3,000 word family level was a minimum for reading unsimplified texts. Similarly, another study of Hirsh and Nation (1992, cited in Cobb and Horst, 2001) also agrees that unsimplified texts could be comprehended when 95% of tokens are known i.e. approximately one unknown word per two printed lines.

Consequently, in current literature, 95% coverage of words known in the text is considered the minimum requirement for reading comprehension. Schmitt (2000, p.152) comments, "A figure of 95% known words crops up in the literature frequently, and at the moment this seems to be a reasonable estimate". Similarly, Nation (2001, p.146) points out, "The safest measure to use in defining the threshold is the coverage (word token) measure which Laufer (1992) found to be around 95%". This means that, in order to adequately understand a piece of text, students need to be familiar with at least 95% of the words occurring in that text. If their lexical knowledge is below this threshold, their ability to comprehend the reading text will unlikely be adequate.

#### **2.2.2.2 Lexical threshold for vocabulary size**

As previously mentioned, the 95% coverage of known words is the minimum requirement for reading comprehension. Some studies have attempted to identify which words and how many words can provide the critical 95% coverage. Laufer (1997) mentions that the minimal comprehension of Israeli university students highly correlated with knowledge of the 3,000 most frequent words of English. The learners below the 3,000-words level did poorly on the reading test regardless of how

high their academic ability was. Therefore, Laufer (1997, p.23) concludes, “The turning point of vocabulary size for reading comprehension is about 3,000 word families”, and the level of 3,000 word families is regarded as a minimum for the reading comprehension of unsimplified texts.

Regarding the classification of word types mentioned earlier, many specialists such as Nation (2001), Cobb and Horst (2001), Coxhead and Nation (2001), and Nation and Waring (1997) indicate that just over 90% of the running words in academic texts can be accounted for by two established words lists i.e. the GSL with either the UWL or the AWL. Adding the GSL to either the UWL or the AWL constitutes a set of about 3,000 academic words shared in all disciplines, and this set also provides about 90% text coverage which is close to the estimated threshold of 95%. As a result, knowledge of the GSL together with either the UWL or the AWL is seen as the lexical knowledge base for reading in any academic domain. Learning these 3,000 word families should be a high priority before students can start to learn at a more advanced level.

The literature on vocabulary types and lexical thresholds for academic reading are consistent in revealing that the two types of high frequency and academic words included in the GSL and the AWL are necessary for being a critical basis for academic reading. In the present study, therefore, it is more appropriate to select words based on these wordlists since the target words are likely to be used frequently by the students in their real academic reading.

### **2.2.3. Assessment of students' vocabulary size**

Apart from identifying and teaching the target words, the measure of students' knowledge of these words is also important. After study, students' vocabulary size should be measured to determine how many of the target words have been learnt. In the present study, the measure of students' vocabulary size or definitional knowledge is designed based on the established vocabulary level tests. Some tests have been developed specifically for assessing students' vocabulary size from the GSL and the UWL/AWL since these two wordlists are regarded as a prerequisite for coping with academic texts. These tests are used to measure two types of knowledge: one is for assessing receptive knowledge and the other is for assessing productive knowledge. The present study is only concerned with the receptive version




of these tests. Widely used tests of receptive knowledge are Nation's (1990), Beglar and Hunt's (1999), and Schmitt, Schmitt and Clapham's (2000) Vocabulary Level Tests whereas those of productive knowledge are Laufer and Nation's (1995) and Laufer and Nation's (1999) Vocabulary Level Tests.

These Vocabulary Level Tests are often recommended (Nation, 2001; Coxhead and Nation, 2001; Beglar, 2000; and Read, 1997) for assessing students' overall vocabulary size because they are well researched and reliable. The Vocabulary Level Tests at 2000 word level from the GSL and Academic Word Level from the UWL/AWL are typically employed to measure the breadth of the learners' vocabulary size as well as to specify what levels of basic lexical knowledge the learners have. These two levels are claimed to represent around 3,000 word families regarded as a lexical critical basis for academic reading since these words frequently occur in various academic text types in all disciplines. According to Nation (2001, p.21), the Vocabulary Level Test "gives credit for partial knowledge of words. Its main purpose is to let teachers quickly find out whether learners need to be working on high-frequency or low-frequency words, and roughly how much work needs to be done on these words". In addition, Beglar (2000, p.2) confirms that the test is classified as a sensitive vocabulary test, which means that the format is sensitive to partial word knowledge. According to him, "This test is designed to estimate examinees' basic knowledge of common word meanings, and specifically, the extent to which they know the common meanings of words at the 2000, 3000, 5000, 10000 and University Word Levels". The Vocabulary Level Tests have been used extensively as diagnostic and placement tests such as at Sultan Qaboos University in Oman (Cobb, 1977) and at Temple University Japan's Corporate Education Program (Beglar, 2000).

One of the well-known Receptive Versions of Vocabulary Level Tests was developed by Paul Nation. This version consists of four equivalent forms of six Word Levels i.e. 1000, 2000, 3000, 5000, 10000 and University Word Levels. The first 1000 Word Level Test consists of 39 questions, each of which has three options for test takers to decide whether a particular question is true, not true, or not understood as in the following examples.

**Instructions:** Write T if a sentence is true. Write N if it is not true. Write X if you do not understand the sentence. The first one has been answered for you.

1. We cut time into minutes, hours and days. T
2. This one is little.  $\rightarrow \triangle$   \_\_\_\_\_
3. Some children call their mother Mama. \_\_\_\_\_
4. *Show me the way to do it* means 'show me how to do it'. \_\_\_\_\_

(Extracted from Nation's (1993) 1000 Word Level Test A published in Nation, 2001, p.412)

The other Levels are in the same format: i.e. 30 items including 10 sets of 3 definitions and 6 words at each level. Recently, Beglar and Hunt (1999) have revised and validated the four forms of the 2000 and University Levels by means of Rasch Item Analyses. Subsequently, they proposed two revised forms, concluding that their content validity is greater than that of the original ones, and both forms are adequately equivalent. Each revised form consists of 27 items making up 9 sets of 3 definitions and 6 words. More recently, another version of Vocabulary Level Tests has been developed by Norbert Schmitt, Diane Schmitt, and C. Clapham (published in Nation, 2001 and Schmitt, 2000). According to Nation (2001), this version includes a pair of equivalent forms and is a major improvement on Nation's original Test, which it replaces. Each form consists of 5 Levels i.e. 2000, 3000, 5000, Academic and 10000 Word Levels. Each Level includes 30 items making up 10 sets of 3 definitions and 6 words. The test-taker's task is to match the words with their definitions, as in the following examples.

**Instructions:** Choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

- |             |                                     |
|-------------|-------------------------------------|
| 1. business |                                     |
| 2. clock    | <u>6</u> part of a house            |
| 3. horse    | <u>3</u> animal with four legs      |
| 4. pencil   | <u>4</u> something used for writing |
| 5. shoe     |                                     |
| 6. wall     |                                     |

- 
- |           |                                     |
|-----------|-------------------------------------|
| 1. copy   |                                     |
| 2. event  | _____ end or highest point          |
| 3. motor  | _____ this moves a car              |
| 4. pity   | _____ thing made to be like another |
| 5. profit |                                     |
| 6. tip    |                                     |

(Extracted from Norbert Schmitt, Diane Schmitt and C. Clapham's 2000 Word Level Test B published in Nation, 2001, p.416)

The way to interpret the result is in percentage. Beglar (2000, p.2) exemplifies, “If learner A scores 9 out of 12 (75%) on the 2000 word level, s/he probably knows approximately 75% (1,500) of the first 2000 words of English. And this logic can be applied to the results of the rest of the tests”. Nation (2001) suggests that a score of at least 25 out of 30 (or over 80%) is desirable for each level.

In the present study, the design format of these vocabulary level tests is adapted to develop the measures of students’ vocabulary size or definitional knowledge. These measures are the Definitional Part in four review tasks as well as in the pretest, the immediate posttest and the delayed posttest.

### **2.3. Vocabulary Acquisition and Retention**

In dealing with vocabulary learning, the present study focuses attention on the measurement of two levels of vocabulary knowledge: definitional knowledge and transferable knowledge. Accordingly, both knowledge types are described before the nature of vocabulary acquisition and retention is reviewed in order to find the best ways which may help learners to learn, retain and retrieve these knowledge types.

#### **2.3.1. Definitional knowledge and transferable knowledge**

The accumulation of vocabulary acquisition is concerned with the breadth and depth of knowledge. According to Qian (1999, p.282), “*breadth of vocabulary knowledge* is defined as vocabulary size or the number of words for which a learner has at least some minimum knowledge of meaning. On the other hand, *depth of vocabulary knowledge* is defined as a learner’s level of knowledge of various aspects of a given word, or how well the learner knows this word”. As mentioned earlier, definitional knowledge of a word is regarded as a shallow level of understanding and likely to occur at early encounters of that word. Nagy (1997, p. 73) mentions, “Definition-based learning typically involves memorizing (or attempting to memorize) brief definitions representing only a single meaning of the word to be learned, and hence lead to only a shallow level of word knowledge”. According to him, definitional knowledge usually occurs at an initial stage to learn new words. However, simply memorizing word definitions cannot guarantee lexical transfer to

other contexts because this type of knowledge is superficial and is unlikely to increase comprehension of texts. It is often found that students know definitions, yet apparently are unable to comprehend textual information.

To increase comprehension of a text, the quality of word knowledge is necessary for lexical knowledge transfer. It is suggested that the ability to transfer lexical knowledge to new contexts demands deep lexical knowledge, and transferable knowledge is possibly enhanced by setting optimum conditions for vocabulary learning. According to Cobb (1997a and b), learning words in various contexts can increase quality of word knowledge since students' ability to transfer their new lexical knowledge to reading comprehension appears. Similarly, Nagy (1997) suggests that the instruction should involve multiple exposures to the words in context, and require deep processing of information about the words. He insists that most lexical knowledge is attributed to encounters with the words in various contexts since no single encounter with a word can lead to any great depth of word knowledge.

### **2.3.2. Incremental nature of vocabulary acquisition and retention**

Vocabulary acquisition and retention possesses *incremental* nature. Knowledge of each word feature needs to be accumulated with knowledge of other features before that word is properly acquired. Meeting a word only once is not sufficient for that word to be learned and retained well since each word possesses more than one feature such as form, pronunciation, meaning, grammatical function, collocation etc. Nation (2001) points out that there is so much to know about each word that one meeting with it is not sufficient to gain all necessary information because vocabulary items must not only be known but they must be known so well that they can be easily accessed. Similarly, Schmitt (2000) mentions that complete mastery of a word entails a number of components of word knowledge, not all of which can be completely learned simultaneously. At each encounter, only one or a few parts of a word can be acquired. However, such knowledge is not a guarantee of word use and retention at all. In order to consolidate word knowledge, the exposure to a particular word should be repeated. The accumulation of knowledge of different aspects of a word can strengthen retention as well as retrieval of such lexical knowledge. The more encounters of a word, the easier that word is likely to be acquired and retained.

Knowledge of vocabulary is accumulated at various levels. Initially, knowing a word definition is likely to take place at the early encounters of a particular word so such knowledge of definition is regarded as a shallow level of knowledge. As knowledge of other features is accumulated and consolidated, knowledge of the word will be increasingly deep. Word knowledge must be deep enough for ensuring lexical knowledge transfer to various language uses. This is concerned with definitional knowledge and lexical knowledge transfer. The incremental nature of vocabulary knowledge is also described by Henricksen (1999, cited in Schmitt, 2000), who identifies lexical knowledge in three dimensions. In the first dimension, learners can have knowledge ranging from zero to partial to precise. In the second dimension, depth of knowledge requires mastery of a number of lexical aspects. The last dimension is receptive and productive mastery. It is often found that a word is learned receptively before it can be used productively.

The incremental nature of vocabulary acquisition and retention has some implications for vocabulary teaching. This means that learners should be provided enough opportunities to repeatedly meet each to-be-learned word in order to consolidate their knowledge. It is very unlikely that a learner is able to grasp even one meaning of a word in one encounter. After the learner meets the words through a variety of activities and in different contexts, a more accurate understanding of its meaning and use will develop. According to Sokmen (1997), re-encountering a new word has significant reward in word retention and long-term memory. Therefore, the repetition of word encounters is important for facilitating learners to acquire lexical knowledge. In classroom practice, to provide multiple exposures of new words to students, the to-be-learned words should be recycled in every possible way. Coady (1997) suggests that proficiency results from a sufficient number of meaningful encounters with the target language. It can be concluded that the more students encounter each word in various contexts, the better they learn and remember it. However, the number of encounters necessary for learning a particular word is quite controversial. The minimum point is acceptable at 5 encounters. Cobb and Horst (2001) accept the idea that stable learning requires meeting a to-be-studied word at least five or eight times, and use this criterion in their development of lexical tutoring computer program to increase students' vocabulary size within limited time.

To summarize, the quality of lexical knowledge seems to take place and be strengthened through meeting the to-be-learned words in a variety of natural contexts several times so that such knowledge can be retained and retrieved for being transferred to new contexts. In other words, the number of word encounters in a variety of contexts is essential in facilitating word retention, retrieval and transfer to new contexts. Therefore, the lessons in the present study are designed according to these principles both in the concordance and non-concordance versions.

## **2.4. Vocabulary Instruction**

In this topic, the background of vocabulary instruction is first highlighted to overview the different roles of vocabulary along various stages of ELT development. Then, the main current approaches to vocabulary instruction are reviewed to determine the strengths of each approach in their compatibility with the implementation of the concordance-based method. This is to enhance the effectiveness of the concordance-based method.

### **2.4.1. Background**

Since the field of English as a second/foreign language (ESL/EFL) was developed as a discipline in the 1950s, there have been numerous different approaches to language teaching/learning, each with different perspectives on vocabulary. Trends in the contexts of ESL/EFL show that the vocabulary component has occupied different statuses in various approaches. Surprisingly, despite being regarded as an important component of language, the historical role of vocabulary in language instruction was frequently subservient to other components, according to Schmitt (2000), O'Dell (1997), and Zimmerman (1997). Most approaches did not explicitly state the methods in dealing with vocabulary. It was assumed that vocabulary learning could happen naturally alongside the other elements of language: grammar, structures, functions, notions, or communication strategies. In the traditional approaches whereby emphasis was shifted between *language analysis* and *language use*, vocabulary was often neglected. In grammar-translation and structural syllabuses, for example, syntax and phonology were given priority whereas language use and communication skills

were predominant in function/notional and communicative syllabuses. In these syllabuses, vocabulary was usually introduced in such a way that suited the presentation of grammar or functions, or through texts used for various structural or communicative purposes.

It was not until the 1980s that attention was paid directly to vocabulary teaching. Schmitt (2000) and O'Dell (1997) mention that the increase of interest in vocabulary results from the influence of modern technology with the advent of computer analysis techniques. Huge language corpora of authentic data can be compiled conveniently. Consequently, it is possible for linguists and lexicographers to conduct extensive and objective studies based on large corpora with far more details than ever. Knowledge derived from corpus-based studies increases an awareness of the importance of vocabulary. Findings from the studies of actual language use have reflected the perceived need for more accurate language description. Some studies have led to considerable interest in the significance of large chunks of language, variously known as *lexical items*, *lexical phrases* and *prefabricated units*. It is even argued that lexical items are central to language use and should be central to language teaching (Zimmerman, 1997).

This changed perception in language description led to immediate effects in the area of English Language Teaching (ELT), and marked the turning points for syllabus designs and pedagogical practice. According to O'Dell (1997), four major new editions of EFL dictionaries were published in 1995 with significant features drawing on the lexical insights provided by massive language corpora. In addition, two tools for syllabus design have been improved with large corpora. One is the compilation of better word frequency lists that allow more confidence in word selection and grading. The other is concordance output that provides an overview of how a particular word is used and practised. In addition, more scholars encourage lexical syllabuses in EFL, which are based on frequency and concordance data.

As corpus-based studies have gained increasing popularity, Schmitt (2000) notices that lexical researches have been conducted in two major strands: the patterning of vocabulary and systematization of word selection. Such research is known as the '*Vocabulary Control Movement*'. According to Schmitt, there were two competing approaches to Vocabulary Movement. The first approach attempted to limit English vocabulary to the minimum necessary for the clear statement of ideas.

One result of this approach was *Basic English* of only 850 words devised for use in regular communication. However, this approach seemed unsuccessful and *Basic English* was perceived as ‘unnatural’ English. On the other hand, the second approach was more convincing with the attempts to use systematic criteria based on frequency information in order to select the most useful words for language learning. Consequently, several lists of vocabulary useful for particular types of reading were developed. An outstanding sample of this approach is Michael West’s (1953) *General Service List of English Words (GSL)* consisting of about 2,000 word families.

The trends over the past two decades have assigned an increasing importance to vocabulary work in teaching ESL/EFL. With the reorientation in language description based on corpus-based studies, the perception of language nature and vocabulary roles has changed remarkably. Schmitt (2000, p.68) mentions, “Insights from corpus research have revolutionized the way we view language, particularly words and their relationship”. Currently, vocabulary occupies an outstanding position in ELT, and is no longer subservient to other components of language learning.

#### **2.4.2 Approaches to vocabulary instruction**

In current practice, the emphases in teaching vocabulary in second or foreign language are based on two extreme approaches. According to Coady’s (1997) extensive review of published research in L2 vocabulary instruction, one extreme is ‘*incidental learning*’ through exposure to language use in contexts whereas the other extreme is ‘*explicit learning*’ through the focused study of words. Between both approaches is ‘*strategy learning*’. Viewed as a continuum based on instruction requirement, incidental learning does not need teaching at all, strategy learning demands some instruction, but explicit learning needs formal instruction. With incidental learning, vocabulary is learned incidentally or naturally from contexts and is not necessarily taught at all. With strategy learning, contextual learning is also valued but some learning strategies need to be taught for effective learning from contexts. In contrast to these two approaches, explicit learning argues for formal instruction of vocabulary by using a combination of techniques. These approaches are discussed in the following sub-sections.



### 2.4.2.1 Incidental learning

*'Incidental learning'* is *'contextual learning'*. It imitates a natural process of L1 acquisition by exposing learners to a variety of contexts when their attention is not on the language itself, but rather on the use of language. Exposure to a variety of contexts is assumed to contribute to the understanding of the depth of the word meaning. Schmitt (2000, p.120) explains, "Incidental learning can occur when one is using language for communicative purposes and so gives a double benefit for time expanded". According to Coady (1997, p.286), the contextual acquisition research does demonstrate that most vocabulary knowledge comes from meaningful language encounters. Learning is more successful if the language is authentic, rich in content, enjoyable, and, above all, comprehensible. To enhance incidental learning, learners have to read a large number of texts or converse for quite some time to come across and acquire particular words.

One major method for enhancing incidental learning is *'extensive reading'*. This method is based on a content-based approach in which students are assigned to do a lot of extra reading. The materials are authentic texts which are not designed specifically for language learning. They may be any books or articles for academic reading in other subject matters or for pleasure reading in leisure time. In this approach, choices are also provided for students to select any texts they want to read. It is assumed that vocabulary is acquired incidentally and naturally when students encounter particular words in rich contexts. Nation (2001) mentions three reasons for its appeal i.e. allowing for learning at one's own level, facilitating a variety of interests and motivation, and making it possible to learn outside classes.

However, the major disadvantages of incidental learning arise from its natural process which is slow and gradual. It takes a long time before successful learning takes place whereas students have limited time for study. In learning language for academic purposes, for example, students cannot learn the necessary skills fast and efficiently enough if they try to adopt this time-consuming approach. In addition, this approach lacks the focused attention for directing students in particular ways. Moreover, learning from contexts will be successful only if the occurrences of target words are incidentally frequent enough. This is not always the case, especially with relatively infrequent words.

#### 2.4.2.2 Strategy learning

'*Strategy learning*' is related to the top-down, naturalistic, and communicative approaches of the 1970s and 1980s. Like the incidental learning approach, context is a major source of vocabulary learning. This approach is expected to compensate for the limitation of incidental learning by focusing on how well students can deal with contexts on their own. It emphasizes teaching specific learning strategies to enable students to learn from contexts effectively. Accordingly, inference skills are perceived as primary strategies to deal with new or unknown words. Learners are taught the strategies of '*inferring from contexts*' by recognizing clues in contexts, using monolingual dictionaries, and not defining words with their bilingual equivalents (Sokmen 1997). Vocabulary instruction is, therefore, implicit. Vocabulary acquisition is assumed to happen mainly through guessing words in contexts.

This approach is very appealing to many scholars so several EFL textbooks based on this approach have been published with the main focus on inferring word meaning from contexts. However, according to Coady (1997), academic reading research indicates that this natural learning cannot provide the literacy skills necessary for EAP students to cope with academic demands. At least five potential problems occur when focusing solely on inference skills. Strategy learning is a slow process and guessing word meaning from contexts is an error-prone process and does not guarantee long term-retention of word knowledge. In addition, it usually causes low proficiency students to become frustrated and ignores the fact that learners have different styles of learning (Sokmen, 1997).

#### 2.4.2.3 Explicit learning

"*Explicit learning* focuses directly on the information to be learned, which gives the greatest chance for its acquisition" (Schmitt, 2000, p.120). In this approach, vocabulary is deliberately taught and students' attention is drawn directly to the lexical items being studied. The explicit instruction of vocabulary is the full attempt to teach certain types of vocabulary by using various teaching methods or techniques. This contrasts sharply with strategy learning and incidental learning which have either implicit or no instruction at all. More and more research studies emphasize the need for explicit vocabulary instruction. They point to the ineffectiveness of just using implicit vocabulary instruction and the need to

accompany it with a much stronger word level or bottom-up approach than had been previously advocated (Sokmen, 1997). In ESL/EFL situations, the environments do not provide rich contexts and the major sources of language are from classrooms. Given a limited time in schools or universities, learning has to be accelerated and learners need formal instruction to prepare themselves to cope with the demand of real language use.

According to Coady (1997), findings from reading and lexicon size research suggest the need for explicit learning/instruction. The minimum requirements of vocabulary or lexical thresholds have been established from such studies. To cope with various kinds of language uses such as academic reading, students must possess a certain size of vocabulary, ranging from 2,000-3,000 word families. The establishment of vocabulary size is so convincing that certain types of vocabulary such as from the GSL, the UWL and the AWL are considered worth teaching. To ensure sufficient encounters of these words, explicit teaching is obviously needed.

With explicit instruction, any teaching techniques can be used, even direct memorization of certain words, if they serve particular learning purposes. For example, extensive reading for incidental learning can be integrated into this approach but in a controlled or simplified way such as glossing texts, using graded readers etc. Inferencing strategies can also be explicitly taught in combination with any other method. Currently, there are many more techniques for teaching vocabulary explicitly. Many studies related to vocabulary instruction have suggested various principles, guidelines, activities and tasks. For example, Nation (2001) recommends three important general processes that may lead to vocabulary acquisition. These processes include '*noticing*', '*retrieval*' and '*creative (generative) use*'. The first process of '*noticing*' is to draw learners' attention to the word as a useful language item. The second process of '*retrieval*' is to arrange activities for new learned words to be subsequently retrieved during the tasks in order that the memory of those words will be strengthened. The last process of '*creative or generative use*' is to provide chances for students to reproduce or use new learned words in subsequent activities. In addition, Coady (1997) indicates three main principles underlying effective teaching i.e. providing both definitional and contextual information, allowing students to process information, and facilitating multiple exposures of each word.

#### 2.4.2.4 Learning vocabulary through reading

The trends in ESL/EFL contexts clearly show that the pendulum of vocabulary instruction has swung back and forth in language acquisition and instruction approaches. As mentioned earlier, the shift is from the direct teaching of vocabulary in the grammar-translation method to the incidental learning in the communicative approach, and now back to a compromise between implicit and explicit teaching. At present, it is accepted that incidental, strategy and explicit learning are all necessary for ESL/EFL learners, despite some limitations of each. All these three current approaches should be seen as distinct but complementary to one another. It is possible to integrate these approaches by '*learning vocabulary through reading*' and this seems to be the best practice at present. According to Schmitt (2000), there is plenty of evidence that learners can acquire vocabulary from reading. Moreover, Coady (1997) notices that related research seems to demonstrate that *systematic vocabulary instruction* together with *learning vocabulary through reading* is a more successful approach than simply learning through contexts alone.

With the method of learning vocabulary through reading, the integration of approaches in explicit learning, strategy learning and implicit learning is possible. Although explicit teaching is most appealing at present, it can cover teaching only some elements of lexical knowledge due to time limitation. Taking an incremental view of vocabulary acquisition, students have to meet a word in different contexts to expand what is known about it. In addition, to consolidate memory of that word, multiple exposures and creative/generative use of a word are needed. Therefore, it is impractical for explicit approach to contextualize all target words or practice all the creative uses of a word for students to totally master them.

To deal with such a problem, strategy learning as well as incidental learning should be promoted to foster students' independent learning. One possible way is to explicitly teach vocabulary through reading at the beginning level before moving to strategy training and finally to incidental learning at higher levels. Findings from lexical threshold research can be used as criteria for determining the boundary of each level. Many scholars such as Nation (2001), Cobb and Horst (2001), Coxhead and Nation (2001) and Nation and Waring (1997) suggest that about 2,000 word families of high frequency words should be properly mastered for general language use and about 3,000 word families for academic purposes. This means that

before students are able to learn useful strategies for guessing word meaning from contexts effectively, about 3,000 word families should be acquired. After these 3,000 word families are explicitly and properly learned, students should be trained to use inference strategies so that they can deal with technical or low frequency words before moving on to incidental learning with extensive reading.

Many techniques have been suggested for the explicit instruction of vocabulary through reading. Nation (2001) and Schmitt (2000), for example, propose that certain words in authentic texts for reading may be made salient, such as by glossing them clearly at the books' margins, or the texts may be simplified. In addition, *intensive reading* of short texts is useful to facilitate text understanding as well as to direct a lot of attention to the vocabulary, grammar and discourse of the texts. With intensive reading, a number of vocabulary and reading exercises must be provided with each reading passage. Moreover, extensive reading is also possible by using *graded readers* with beginning students, *narrow reading* with intermediate students, and a wide variety of authentic texts with advanced students. *Graded readers* are authentic books which are graded according to levels of readability whereas *narrow reading* means reading numerous authentic texts, but all on the same topic in order that much of the topic-specific vocabulary will be repeated throughout the course of reading. Schmitt (2000) emphasizes the benefit of narrow reading in that it can accelerate access to authentic materials.

## 2.5 A Concordance-based Method

A '*concordance-based method*' is the method adapted from a corpus technique widely used for linguistic analysis in the fields of computational linguistics and lexicography. This method essentially involves corpus compilation from authentic texts and a concordancing program for accessing a corpus and then producing concordance output. Since a corpus and a concordancer are always used together in this method of language analysis, the terms '*corpus-based method*' and '*concordance-based method*' usually co-occur in related literature so they are often interchangeable in most cases. When language corpora were introduced to language instruction a few decades ago, the concordancer was also exploited in the pedagogical field as an

indispensable corpus tool. In language pedagogy, the application of this method is underpinned by a learning approach called '*data-driven learning (DDL)*'. *DDL* is sometimes referred to in the literature as *classroom concordancing*, although a slight distinction can be made between the two terms. According to Sripicharn (2000), *DDL* refers to the methodological framework of the approach whereas *classroom concordancing* refers to the practical aspect of the approach. Thus, these two terms are used interchangeably in his Ph.D. thesis. Similarly, in this paper, these terms are mostly used interchangeably.

### 2.5.1 Background

Language corpora have long been exploited for language study. They were undertaken manually before computers were available. As technology advancement has increased the power and capacity of computers, corpora have increased dramatically in size, variety and ease of access. Simultaneously, an expanding range of software has been developed to process corpora and to access the information they contain. With the rapid advancement of computational linguistics, the computer-based corpora have led to a new discipline known as *corpus linguistics* since the last few decades (Kennedy, 1998). The field of study is based on bodies of texts as the domain of study and as the source of evidence for linguistic description and argumentation. Work related to corpus linguistics is being done in various fields and is multiplying at a very fast rate.

Considerable corpus-based work has been increasingly developed especially in the field of lexicography. Large-scale corpora have been exploited to investigate language as it is actually used. These corpora have dramatically improved the quality of reference materials. These references analyze and report precisely and confidently how language is actually used rather than providing prescribed information. Resulting from such corpus-based work, English descriptive grammar is reassessed as evident in the publications of the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad and Finegan 1999) and *An Empirical Grammar of the English Verb: Modal Verbs* (Mindt, 1995). Modern English dictionaries published currently all indicate that they are based on findings from corpus-based studies. These dictionaries, for example, are the *Longman Dictionary of Contemporary English* (2005) based on the Longman Corpus Network,

the *Cambridge International Dictionary of English* (1995) based on the Cambridge Language Survey corpus, and the *Collins COBUILD English Language Dictionary* (1995) based on the Collins Cobuild Database. The number of corpus-related studies is good evidence of the growing interest in corpus-based research. Thus, the number of corpora has mushroomed considerably and they have become more widely accessible. At present, an electronic corpus has become a universal resource for most linguistic investigation.

However, the area of English Language Teaching (ELT) has been rather slow to incorporate corpus-based method into its working practice, compared to other related fields of study. Despite being introduced to ELT contexts in the 1980s, the application of concordances in EFL classroom in the 1990s was still in its infancy as a language teaching technique (Stevens, 1995; and Fox, 1998). At the initial stage of its arrival in ELT, the method was exploited exclusively by developers of curricula, syllabuses and materials in order to determine the representative language of their target language use. Later, language teachers were encouraged to exploit corpora as the linguistic informant to update their linguistic knowledge with current language use, and as a source of input for preparing classroom materials and for searching authentic linguistic examples. Recently the method has been used in language classrooms not only for materials preparation but also for language learning. Learners are given more opportunities to have direct contact with relevant authentic information in corpora. This has led to the emergence of a learning approach called '*Data-driven Learning*' (DDL), in which students are assigned to work with raw information taken directly from corpora. It is based on the assumption that students can acquire language effectively when they engage in language analysis. The method can draw students' active involvement in the learning process by encouraging them to observe linguistic input, form hypotheses and draw their own conclusions about word/phrase meaning and grammar rules based on the examination of authentic linguistic evidence. Accordingly, learning and self-discovery possibly take place when the students are placed metaphorically in the position of researchers.

Currently, the role of language corpora in language teaching has gradually become prominent. Since corpora and concordancing programs have become available and more easily accessible for teachers and learners, their very potential application has been seen to offer new and exciting directions in developing

curricula, syllabuses and teaching materials as well as facilitating students to make direct discoveries about language. The increase in the number of published works in ELT is good evidence of its rapid growth. Several papers such as in Gavioli and Aston (2001), Conrad (2000), Fox (1998) and Owen (1997) discuss the important roles of corpora in classroom pedagogy. It is now established that a basic corpus technique plays a major role in shaping pedagogical practice.

### 2.5.2 Description of the concordance-based method

The concordance-based method is a method of language analysis for linguistic study. It consists of three main components i.e. a corpus, a concordancer and a concordance. In language analysis, a *corpus* is like a database, a *concordancer* is a corpus-accessing tool working like a search engine for searching linguistic information of words or phrases to be studied, and a *concordance* is a formatted display where all occurrences of any particular word are listed together in the contexts.

#### 2.5.2.1 A corpus

A corpus is a collection of texts compiled for linguistic study. The term '*corpus*' comes from the Latin word for '*body*' and it has retained this meaning i.e. '*any body of text*' (McEnery and Wilson, 2001). However, in the context of linguistic study, this simple definition is considered insufficient because a corpus cannot be seen as just a collection of texts but it should be gathered on a linguistic basis. The definition of a corpus as '*any body of text*' may lead to confusion between the term '*corpus*' and '*archive*' so a distinction between them is made. Accordingly, a *corpus* is generally referred to as a collection of texts gathered according to particular principles for some particular purposes whereas an *archive* refers to a collection where various kinds of texts are stored simply because each individual text is interesting in itself.

Crystal (1994, p.410) stated that a *corpus* is 'a representative example of language, compiled for the purpose of linguistic analyses'. In his 1991 work, he also defines a corpus as "a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or a means of verifying hypotheses about a language". Leech (1997)



pointed out two consecutive descriptions. According to him, linguists have traditionally used the term *corpus* to designate a body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research. This body of data may consist of written texts, spoken discourses, or samples of spoken and/or written language. Later, the term *corpus* has been increasingly applied to a body of language material which exists in electronic form and which may be processed by computer for various purposes such as linguistic research and language engineering. Nevertheless, Kennedy (1998) explains that corpora are not necessarily stored electronically so that they can be machine-readable although this is nowadays the norm. According to him, corpus linguistics did not begin with the development of computers and some of the most revealing insights into language use have come from the blend of manual and computer analysis. Kennedy's (1998, p.1) brief definition of a *corpus* is "A body of written text or transcribed speech which can serve as a basis for linguistic analysis and description." Corpora may consist of whole texts or collections of whole texts. They may consist of continuous text samples taken from whole texts or even collection of citations.

Although there are a large number of corpora increasingly available at present, they will not always serve the need of every potential user. Some researchers, material developers, teachers or even students may need to compile their own corpora for particular purposes. In doing so, the criteria can vary from one to another. Prior to compiling a corpus, therefore, the objectives of the study must be clearly set and particular considerations must be taken to ensure the quality of a corpus. Such quality is mainly concerned with the issues of '*representativeness*' of the target language or the size of the corpus. "A corpus is *representative* in the sense that findings based on an analysis of it can be generalized to the language as whole or a specified part of it", according to Leech (1991, cited in Kennedy, 1998, p.62). Pearson (1998) suggests that a corpus must be as big as possible to carry out linguistic studies on language as a whole. In addition, to study on a subset of the language, the representative of the subset in question is another important factor in a corpus compilation.

Typically, most corpora are deliberately designed in a size as big as possible although in some cases the size is not necessarily the most important criteria. For purpose-built corpora in language learning, it is not always necessary to

compile corpora as large as the general purpose ones. In such corpora, a large size is less important because the adequacy of a corpus depends on the intended application. Teachers and learners have rather different objectives from professional linguists so that a small corpus with less systematic analyses may still be sufficiently useful. An enormous size of a corpus may be too large for any practical handling of the students. Aston (1998, p.226) suggests, “small specific corpora have obvious virtues in highlighting recurrent specialised features, but only larger and more general ones seem able to capture less specialised ones, and to contextualise such features against a broader spectrum of abilities and awareness”. According to him, analyzing data in a small specialized corpus potentially allows teachers and learners to contextualize uses encountered against a broader linguistic background.

In compiling one’s own corpus, Aston (2002, p.14) suggests that the web is one excellent resource although complex searches and considerable adaptation are needed. According to him, dividing a corpus into sub-corpora is also an attractive strategy since a small size of a sub-corpus is more manageable and available for being selected according to the desired proportions. In addition, a small specialized corpus can offer a number of practical advantages over a large mixed one because it is relatively simple to compile, analyze, interpret and be familiar with (Aston 1997b and 2001). Accordingly, incidental learning of vocabulary is likely to be less dispersive since linguistic input is confined to specific text-types and more immediate to learners. In determining a corpus size, consideration should be taken as to whether a corpus is sufficient for serving the purpose of the study and for being representative of the target language (Aston, 2001). Many researchers tend to agree that smaller corpora can suffice and appropriate in cases that the investigated phenomena appear with sufficient frequency to provide adequate result. Most papers published in Aston, (ed.) (2001) use relatively small specialized corpora for language learning, ranging from 2,000 to 1,000,000 running words. Some of them are corpora of newspaper articles, transcribed speech, academic writing and classified advertisement.

#### **2.5.2.2 A concordancer and a concordance**

Most corpora are incredibly large and it is a formidable task to study corpus information without the help of a computer. An important tool for

working with language corpora is a '*concordancer*', which is a computer program used to search, access and analyze corpus information, and then to display the output in concordance lines. A *concordance* is "an alphabetical listing of words in a text or collection of texts, together with the contexts in which they appear" (Godwin-Jones, 2001, p.8). In other words, a *concordance* is a list of occurrences of either a particular word, a part of a word, or a combination of words in context. An occurrence of a particular word is usually called a *keyword*.

A typical concordancer allows us to enter a word or phrase and search for multiple examples of how that word is used in speech and writing. More complex concordancers can help us to extract examples from very particular contexts and even discriminate between spoken and written language. With the use of a concordancer to access corpus information, concordances can be produced in a number of formats. The most useful form is a *Keyword in Context (KWIC)* format. A typical *KWIC format* displays the keyword in the center of the line with more contexts on each side of the keyword and each occurrence of the word is listed on a separate line. It is also possible to display the sequences of contexts either on the left or right of the keywords. Therefore, it is convenient to get a picture of the environments where a keyword occurs in a corpus.

Since a concordancer is capable of making a concordance list showing the contexts of every occurrence of a selected word or phrase in a text corpus, it is sometimes called a '*super-index*'. However, most concordancers are more capable than simply indexing words into lines. It is particularly useful in exploring the relationships between words, and it can provide very accurate information about the way language is authentically used. Sorted concordances can provide information on collocation patterning as well as reveal different senses of a word type. Moreover, the relative frequencies of different uses of a word type can be calculated.

### **2.5.2.3 Basic functions of a concordancer**

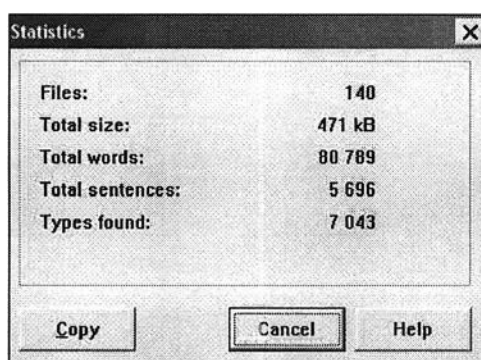
As corpus-based studies are becoming increasingly prominent, a wide range of concordancing programs have been continuously developed for operating sophisticated functions as well as for making them much more user-friendly. Therefore, the operation of different concordancers varies from system to system. Nevertheless, all fundamental principles are still common to all different

concordancers, according to Tribble and Jones (1990). Some basic functions of most concordancers are described below. All examples illustrated below are from the use of *WCONCORD*, a concordancer developed by Martinek and Siegrist in 1999.

### 2.5.2.3.1 Displaying statistical information of a corpus

Most concordancers are very capable of counting words, word types and sentences in the corpus and then showing the statistical information of the corpus as illustrated in Figure 2.1. This information provides the overall idea about the size of a corpus and the total number of words, word types, files and sentences in that corpus. It is helpful in determining the size and reliability of the corpus.

Figure 2.1: Statistical information of a corpus



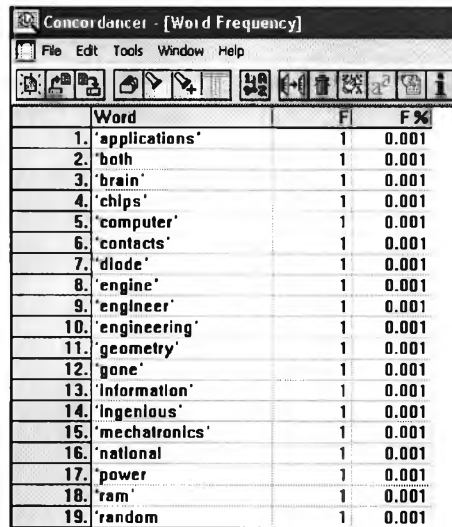
The image shows a window titled 'Statistics' with a close button (X) in the top right corner. Inside the window, there is a table of statistics. At the bottom of the window, there are three buttons: 'Copy', 'Cancel', and 'Help'.

|                  |        |
|------------------|--------|
| Files:           | 140    |
| Total size:      | 471 kB |
| Total words:     | 80 789 |
| Total sentences: | 5 696  |
| Types found:     | 7 043  |

### 2.5.2.3.2 Building word frequency lists

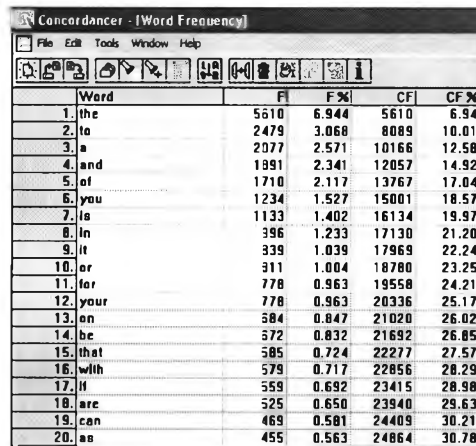
With the capability of word counting, a concordancer can quickly build word frequency lists including all words occurring in a given corpus. When the wordlist is built, the resulting words are usually displayed alphabetically as in Figure 2.2. Apart from a word sort in an alphabetical order, the list can optionally be sorted either in a frequency order or in a retrograde order as illustrated in Figures 2.3 and 2.4. With frequency sort, words are displayed in a descending order ranging from the most frequently occurring words in the corpus to the least frequently occurring ones. When the retrograde order is selected, words are sorted alphabetically according to the endings of words. This type of data sorting is useful for studying the recurrent patterns of word suffixes.

Figure 2.2: A word frequency list sorted by alphabetical order



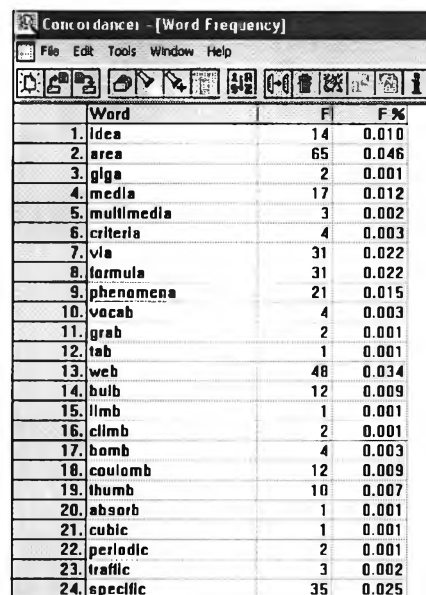
|     | Word           | F | F%    |
|-----|----------------|---|-------|
| 1.  | 'applications' | 1 | 0.001 |
| 2.  | 'both'         | 1 | 0.001 |
| 3.  | 'brain'        | 1 | 0.001 |
| 4.  | 'chips'        | 1 | 0.001 |
| 5.  | 'computer'     | 1 | 0.001 |
| 6.  | 'contacts'     | 1 | 0.001 |
| 7.  | 'diode'        | 1 | 0.001 |
| 8.  | 'engine'       | 1 | 0.001 |
| 9.  | 'engineer'     | 1 | 0.001 |
| 10. | 'engineering'  | 1 | 0.001 |
| 11. | 'geometry'     | 1 | 0.001 |
| 12. | 'gone'         | 1 | 0.001 |
| 13. | 'informaton'   | 1 | 0.001 |
| 14. | 'ingenious'    | 1 | 0.001 |
| 15. | 'mechatronics' | 1 | 0.001 |
| 16. | 'national'     | 1 | 0.001 |
| 17. | 'power'        | 1 | 0.001 |
| 18. | 'ram'          | 1 | 0.001 |
| 19. | 'random'       | 1 | 0.001 |

Figure 2.3: A word frequency list sorted by word-frequency order



|     | Word | F    | F%    | CF    | CF%   |
|-----|------|------|-------|-------|-------|
| 1.  | the  | 5610 | 6.944 | 5610  | 6.94  |
| 2.  | to   | 2479 | 3.068 | 8089  | 10.01 |
| 3.  | a    | 2077 | 2.571 | 10166 | 12.58 |
| 4.  | and  | 1891 | 2.341 | 12057 | 14.92 |
| 5.  | of   | 1710 | 2.117 | 13767 | 17.04 |
| 6.  | you  | 1234 | 1.527 | 15001 | 18.57 |
| 7.  | is   | 1133 | 1.402 | 16134 | 19.97 |
| 8.  | in   | 396  | 1.233 | 17130 | 21.20 |
| 9.  | it   | 339  | 1.039 | 17969 | 22.24 |
| 10. | or   | 311  | 1.004 | 18780 | 23.25 |
| 11. | for  | 778  | 0.963 | 19558 | 24.21 |
| 12. | your | 778  | 0.963 | 20336 | 25.17 |
| 13. | on   | 584  | 0.847 | 21020 | 26.02 |
| 14. | be   | 572  | 0.832 | 21692 | 26.85 |
| 15. | that | 585  | 0.724 | 22277 | 27.57 |
| 16. | with | 579  | 0.717 | 22856 | 28.29 |
| 17. | if   | 559  | 0.692 | 23415 | 28.98 |
| 18. | are  | 525  | 0.650 | 23940 | 29.63 |
| 19. | can  | 469  | 0.581 | 24409 | 30.21 |
| 20. | as   | 455  | 0.563 | 24864 | 30.78 |

Figure 2.4: A word frequency list sorted by retrograde order



|     | Word       | F  | F%    |
|-----|------------|----|-------|
| 1.  | idea       | 14 | 0.010 |
| 2.  | area       | 65 | 0.046 |
| 3.  | glga       | 2  | 0.001 |
| 4.  | media      | 17 | 0.012 |
| 5.  | multimedia | 3  | 0.002 |
| 6.  | criteria   | 4  | 0.003 |
| 7.  | via        | 31 | 0.022 |
| 8.  | formula    | 31 | 0.022 |
| 9.  | phenomena  | 21 | 0.015 |
| 10. | vocab      | 4  | 0.003 |
| 11. | grab       | 2  | 0.001 |
| 12. | tab        | 1  | 0.001 |
| 13. | web        | 48 | 0.034 |
| 14. | bulb       | 12 | 0.009 |
| 15. | llmb       | 1  | 0.001 |
| 16. | climb      | 2  | 0.001 |
| 17. | bomb       | 4  | 0.003 |
| 18. | coulomb    | 12 | 0.009 |
| 19. | thumb      | 10 | 0.007 |
| 20. | absorb     | 1  | 0.001 |
| 21. | cubic      | 1  | 0.001 |
| 22. | periodic   | 2  | 0.001 |
| 23. | traffic    | 3  | 0.002 |
| 24. | specific   | 35 | 0.025 |

Frequency information is very helpful in word selection. By doing a frequency count, it is possible to find out the relative frequency of a word ranging from the most to the least frequent words. It is likely that the most frequent words are selected for study although this is not always the case since function words such as articles and prepositions are usually found in the top ranks of most corpora. For lexical words, however, high frequency words are most highly considered as being worth studying since they are more likely to be found in other contexts. It is also possible to see the frequent use among particular words such as synonyms and spelling variants. For example, a group of near synonyms e.g. 'start', 'begin', and 'commence' can be studied to see which one is used more/less frequently in an informal/formal situation or in a written/spoken mode. Tribble and Jones (1990) recommend that creating wordlists and frequency tables is often the best way to start. With frequency information, it is possible to have a much better idea of which particular items should be properly selected for further studying other aspects of those items. In addition, to set a cut-off point between maximum and minimum frequencies can prevent an overwhelming amount of data. Therefore, it saves a great deal of guesswork if we begin with wordlists and then move on to other types of analyses.

### 2.5.2.3.3 Searching words

A *concordancer* is also capable of searching individual words, phrases and parts/combinations of words. Basically, after a particular to-be-studied word is typed into the program, a concordancer will compile a concordance list according to the occurrences of that word found in the corpus. More than one word can be searched for at a time and the search for collocations or groups of words is also possible. Figure 2.5 illustrates the result from searching the collocation of 'on the other hand'.

Figure 2.5: The concordance output of searching 'on the other hand'

The screenshot shows a window titled 'Concordance - [Concordance: 1\*]' with a menu bar (File, Edit, Tools, Window, Help) and a toolbar. Below the toolbar, the search term 'On the other (hand) . velocity is direction-aware.' is displayed. The main area contains a table of results:

| Line | Text  | Word         | Text   |
|------|---|--------------|--|
| 1.   |   | On the other | hand . velocity is direction-aware.                        |
| 2.   |   | On the other | hand . If you are accelerating upward in an elevator, the  |
| 3.   |   | On the other | hand . the keyboard is an example of an input device wh    |
| 4.   |   | On the other | hand . If the cells are connected in parallel, the voltage |
| 5.   |   | On the other | hand . If the cells have been used for some time, they n   |
| 6.   | The Earth's gravitational pull, on the other                          | hand         | . decreases as you move farther away from the E            |
| 7.   | The lamp filament, on the other                                       | hand         | . is made up of very thin wire.                            |
| 8.   | Potential energy, on the other  | hand         | . is stored energy.  |
| 9.   | and better individual access to cultural diversity, with on the other | hand         | . some reduction in diversity through assimilation,        |
| 10.  | Output devices on the other   | hand         | . decode the data into information which can be un         |

A wildcard search of most concordancers can make the searches more specific and effective. This allows searching for a 'root' word with a 'wildcard character' i.e. a symbol standing for one or more unspecified characters. Different symbols are used for the 'wildcard character' depending on the concordancing programs. Typically, the question mark (?) is a substitute for a wildcard character for any single character and the asterisk (\*) is a substitute for a wildcard character for any zero or more characters. These symbols can be put at the beginning of words, at the end of words, or in the middle of words. For example, a search for 't??k' may find *talk, tank, task, took* etc whereas a search for 't\*k' may result in *talk, tank, task, teamwork, thank, thick, think, took, track, trademark, truck* etc. Another example is from the search of '?ing' which may find the word with one more character in front of '-ing' such as *king, ring* etc whereas the search of '\*ing' may find all corpus words with the '-ing' endings. Similarly, the search of 'depend?' may result in *depends* whereas the search of '\*depend\*' may find *independent, independence, depends, depended, depending, dependence, dependent* etc. as in Figure 2.6.

Figure 2.6: The wildcard search of '\*depend\*' sorted by left contexts

| Line | Left Context  | Word           | Right Context   |
|------|---|----------------|---|
| 1.   | Magnetic forces will also                                       | depend         | on the velocities of the two objects.                       |
| 2.   | Operation All internal combustion engines                       | depend         | on the chemical process of combustion and explosion.        |
| 3.   | Thermocouple meters Meters that                                 | depend         | on the heating effect of an electric current are used       |
| 4.   | The importance of   | dependable     | electricity generation, transmission and distribution       |
| 5.   | Electric forces between two objects                             | depend         | only on the charges of the two objects and their            |
| 6.   | If a physical result  | depend         | on the right-handed rule that would constitute violation    |
| 7.   | All other forces encountered only                               | depend         | on the relative distance.                                   |
| 8.   | It immediately brings up the question of frame                  | dependence     | of reference. "How can a force depend on the velocity, when |
| 9.   | is to expand in response to environmental concerns, and as our  | dependence     | of reference becomes better understood                      |
| 10.  | edance to DC (theoretically zero), and a higher impedance to AC | dependent      | on the value of inductance and the frequency.               |
| 11.  | Because of Ohm's law, electrical energy losses are              | dependent      | on current flow, not on energy flow.                        |
| 12.  | The series RC circuit also exhibits frequency                   | dependent      | behaviour, but at DC the impedance is infinite (due         |
| 13.  |   | Depending      | on the resistance of the material making up the body        |
| 14.  | Current can be AC or DC, positive or negative.                  | depending      | upon the reference.   |
| 15.  | Different bits can be used                                      | depending      | on the material and type of cut.                            |
| 16.  | The size of the current   | depends        | on the number of electrons passing per second.              |
| 17.  | AC motor speed primarily  | depends        | on the frequency of the AC supply and the amount            |
| 18.  | DC motor speed generally  | depends        | on a combination of the voltage and current flow            |
| 19.  | The strong force is   | independent    | of electric charge, and holds together, for example         |
| 20.  | It is important to understand that mass is                      | independent    | of your position in space.                                  |
| 21.  | If time t is the only   | independent    | variable the dynamic system will be described by            |
| 22.  | The first integrated circuits were developed                    | independently  | by two scientists: Jack Kilby of Texas Instrument           |
| 23.  | In modern cars the front wheels are                             | independently  | suspended from the frame in a manner that permits           |
| 24.  | Within the rotor is the eccentric shaft that turns              | independently  | of both the rotor and the fixed gear.                       |
| 25.  | In a progressively  | interdependent | world where culture tempers and inflames political          |

#### 2.5.2.3.4 Sorting concordance lists

When a concordance list is built, it is not sorted. The resulting list is normally displayed according to the order in which the program finds each word as in the output of searching 'concerned' illustrated in Figure 2.7.

Figure 2.7: The unsorted output of searching 'concerned'

| Line | Context  | Keyword   | Context   |
|------|--|-----------|---|
|      | Both are [concerned] with generating, transferring, and utilizing electrical energy. |           |   |
| 1    | Both are   | concerned | with generating, transferring, and utilizing electric |
| 2    | The chief difference is that electricity is  | concerned | with using that electrical energy in power applica    |
| 3    | lower applications for heat, light, and motors while electronics is                  | concerned | with power control and communications applicati       |
| 4    | Technology is not as   | concerned | about why as it is how                                |
| 5    | Differential calculus Main article derivative Differential calculus is               | concerned | with finding the instantaneous rate of change (or d   |
| 6    | It is  | concerned | with moving packages from one address to anoth        |
| 7    | The operating system is a complex collection of many programs                        | concerned | with keeping the hardware and software compone        |
| 8    | Mathematics and physics, dynamics is the branch of mechanics that is                 | concerned | with the effects of forces on the motion of objects,  |
| 9    | Energy exchange matter or energy, classical thermodynamics is not                    | concerned | with the rate at which such processes take place,     |
| 10   | Because thermodynamics is not  | concerned | with the concept of time, it has been suggested th    |
| 11   | Electrical Safety In electronics we must be  | concerned | with the protection of our equipment from damage      |
| 12   | , George Bernard Shaw Engineering is   | concerned | with the implementation of a solution to a practica   |
| 13   | Self-contained and can still be used as a 2D system without being                    | concerned | with its 3D features.                                 |
| 14   | (Engineering) Industrial Engineering Industrial engineering is                       | concerned | with the design, improvement, and installation of     |
| 15   | Many ordinary Internet users are less  | concerned | about the actual copyright itself but more about th   |
| 16   | Two states that electronic devices in computers can take up are                      | concerned | with voltage levels.                                  |
| 17   | A mathematician is   | concerned | with the exact definition of dy/dx.                   |

Most concordancers allow the list to be sorted to make word observation more convenient. The concordances can be optionally sorted either by the left or the right contexts of the keywords. With the left sort, the first words on the left of the keyword are ordered alphabetically as in Figure 2.8. Similarly, if the right sort is selected, the first words on the right of the keywords are sorted alphabetically as in Figure 2.9.

Figure 2.8: The concordance list sorted by the left contexts

| Line | Context  | Keyword   | Context  |
|------|--|-----------|--|
|      | Electrical Safety In electronics we must be [concerned] with the protection of our equipment from damage and ourselves from electrical shock or worse. |           |  |
| 1    | Electrical Safety In electronics we must be  | concerned | with the protection of our equipment from damage and oursel    |
| 2    | Both are   | concerned | with generating, transferring, and utilizing electrical energy |
| 3    | Two states that electronic devices in computers can take up are  | concerned | with voltage levels.   |
| 4    | Self-contained and can still be used as a 2D system without being  | concerned | with its 3D features.  |
| 5    | Technology is not as   | concerned | about why as it is how.  |
| 6    | (Engineering) Industrial Engineering Industrial engineering is   | concerned | with the design, improvement, and installation of integrate    |
| 7    | , George Bernard Shaw Engineering is   | concerned | with the implementation of a solution to a practical problem   |
| 8    | A mathematician is   | concerned | with the exact definition of dy/dx.                            |
| 9    | Lower applications for heat, light, and motors while electronics is  | concerned | with power control and communications applications for h       |
| 10   | Differential calculus Main article derivative Differential calculus is   | concerned | with finding the instantaneous rate of change (or derivative)  |
| 11   | Mathematics and physics, dynamics is the branch of mechanics that is   | concerned | with the effects of forces on the motion of objects.           |
| 12   | It is  | concerned | with moving packages from one address to another, without      |
| 13   | The chief difference is that electricity is  | concerned | with using that electrical energy in power applications for h  |
| 14   | The operating system is a complex collection of many programs  | concerned | with keeping the hardware and software components of a syst    |
| 15   | Many ordinary Internet users are less  | concerned | about the actual copyright itself but more about the effect of |
| 16   | Because thermodynamics is not  | concerned | with the concept of time, it has been suggested that a helio   |
| 17   | Energy exchange matter or energy, classical thermodynamics is not  | concerned | with the rate at which such processes take place, termed ki    |

Figure 2.9: The concordance list sorted by the right contexts

| Line | Context   | Keyword   | Context   |
|------|---|-----------|---|
|      | Many ordinary Internet users are less [concerned] about the actual copyright itself but more about the effect on the Internet as a whole if lighter copyright result from the infringement. |           |   |
| 1    | Many ordinary Internet users are less   | concerned | about the actual copyright itself but more about th   |
| 2    | Technology is not as  | concerned | about why as it is how.                               |
| 3    | Differential calculus Main article derivative Differential calculus is  | concerned | with finding the instantaneous rate of change (or d   |
| 4    | Both are  | concerned | with generating, transferring, and utilizing electric |
| 5    | Self-contained and can still be used as a 2D system without being   | concerned | with its 3D features.                                 |
| 6    | The operating system is a complex collection of many programs   | concerned | with keeping the hardware and software compone        |
| 7    | It is   | concerned | with moving packages from one address to anothe       |
| 8    | Lower applications for heat, light, and motors while electronics is   | concerned | with power control and communications applicati       |
| 9    | Because thermodynamics is not   | concerned | with the concept of time, it has been suggested th    |
| 10   | (Engineering) Industrial Engineering Industrial engineering is  | concerned | with the design, improvement, and installation of     |
| 11   | Mathematics and physics, dynamics is the branch of mechanics that is  | concerned | with the effects of forces on the motion of objects   |
| 12   | A mathematician is  | concerned | with the exact definition of dy/dx.                   |
| 13   | , George Bernard Shaw Engineering is  | concerned | with the implementation of a solution to a practica   |
| 14   | Electrical Safety In electronics we must be   | concerned | with the protection of our equipment from damage      |
| 15   | Energy exchange matter or energy, classical thermodynamics is not   | concerned | with the rate at which such processes take place,     |
| 16   | The chief difference is that electricity is   | concerned | with using that electrical energy in power applica    |
| 17   | Two states that electronic devices in computers can take up are   | concerned | with voltage levels.                                  |



Sorting words makes it easier to find occurrences that are typical of a particular word such as grammatical information and collocations since possible recurrent patterns of words become more noticeable. For example, if a word '*concerned*' is the target word of study, sorting words is helpful for highlighting possible recurrent patterns. The left sort in Figure 2.8 is helpful for comparing the frequent use of '*concerned*' as a component of a passive form or as a modifier. In addition, students' attention may be drawn to the components of the passive form in each concordance. On the other hand, the right sort in Figure 2.9 is helpful in identifying the typical co-occurring words or prepositions. With the left and right sort, the most frequent collocation of '*concerned*' i.e. '*be concerned with*' is easily identified. Therefore, word sort is very helpful for the study of significant collocations because typical co-occurring words can also be identified from the amount of context in which a keyword appears in concordances.

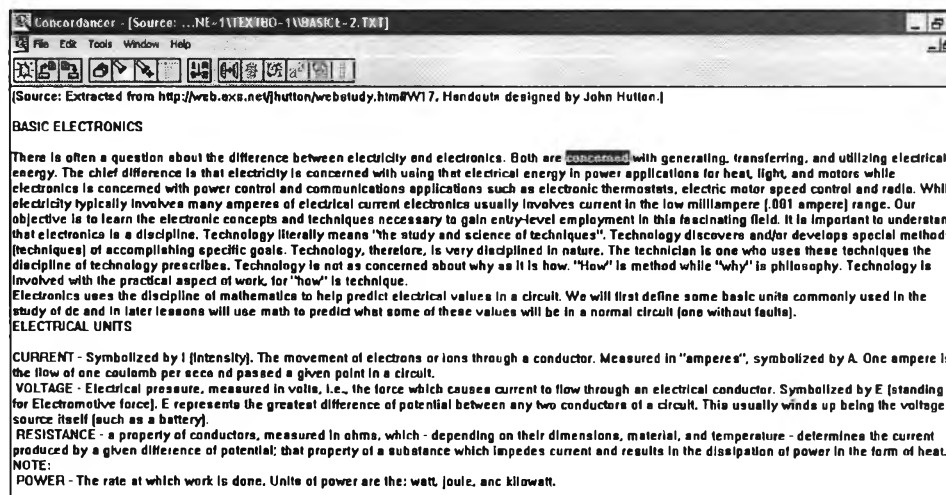
### 2.5.2.3.5 Providing more contexts and word information

Most concordances are usually displayed in fragments rather than in complete sentences. Many concordancers allow the possibility of referring to the source texts to get more context and information about particular words if needed. With WCONCORD, for example, a full sentence of a selected concordance is displayed at the top of the concordance list whereas the computer file name is displayed at the end of each concordance as in Figure 2.10. In addition, the source text of the selected concordance can also be referred to. Figure 2.11 illustrates the source texts in which the keyword is highlighted.

Figure 2.10: The full sentence and the file name of the selected concordance

| Concordancer - [Concordance: concerned]                                      |   |               |  |
|--|---|---------------|--|
| File Edit Tools Window Help  |   |               |  |
| Both are [concerned] with generating, transferring, and utilizing electrical |   |               |  |
|  |   | Full sentence | File names   |
| 1.   | Both are  | concerned     | with generating, transferring, and utilizing electrical en |
| 2.   | The chief difference is that electricity is                   | concerned     | with using that electrical energy in power applications    |
| 3.   | Applications for heat, light, and motors while electronics is | concerned     | with power control and communications applications         |
| 4.   | Technology is not as  | concerned     | about why as it is how.                                    |
| 5.   | calculus Main article derivative Differential calculus is     | concerned     | with finding the instantaneous rate of change (or deriv    |
| 6.   | It is   | concerned     | with moving packages from one address to another, w        |
| 7.   | ing system is a complex collection of many programs           | concerned     | with keeping the hardware and software components of       |
| 8.   | d physics, dynamics is the branch of mechanics that is        | concerned     | with the effects of forces on the motion of objects.       |
| 9.   | ange matter or energy, classical thermodynamics is not        | concerned     | with the rate at which such processes take place, termi    |
| 10.  | Because thermodynamics is not                                 | concerned     | with the concept of time, it has been suggested that a t   |
| 11.  | Electrical Safety In electronics we must be                   | concerned     | with the protection of out equipment from damage and       |
| 12.  | , George Bernard Shaw Engineering is                          | concerned     | with the implementation of a solution to a practical pro   |
| 13.  | ned and can still be used as a 2D system without being        | concerned     | with its 3D features.                                      |
| 14.  | ngineering) Industrial Engineering Industrial engineering is  | concerned     | with the design, improvement, and installation of Integ    |
| 15.  | Many ordinary Internet users are less                         | concerned     | about the actual copyright itself but more about the effe  |
| 16.  | es that electronic devices in computers can take up are       | concerned     | with voltage levels.                                       |
| 17.  | A mathematician is  | concerned     | with the exact definition of dy/dx.                        |

Figure 2.11: The source text of the selected concordance



### 2.5.2.3.6 Manipulating output

Most concordancers also allow the concordance output to be manipulated for various purposes, especially for the preparation of class materials. Among these functions, some duplicate sentences or resulting concordances can be conveniently deleted. The output can also be copied and pasted in order to be converted into normal word-processing programs such as Microsoft Word or Microsoft Word Excel. This makes the practice of materials preparation convenient since linguistic samples can be easily transferred into typical paper-based materials.

### 2.5.3. A learning approach: Data-driven Learning (DDL)

'*Data-driven Learning (DDL)*' is a learning approach inherently associated with classroom concordancing. This approach was initiated by Tim John at the University of Birmingham. According to Murison-Bowie (1996, p.190), DDL is based on the concept originating from John's statement '*research is too serious to be left to the researcher*'. His statement is commonly quoted in many published papers such as in Leech (1997) and Gavioli (1997). According to this statement, John (1991, p.2) considers language learners as being essentially "a research worker whose learning needs to be driven by access to linguistic data". This means that in DDL the students are put in a position similar to that of a researcher in order to examine particular data through the observation of the corpus and then make sense of the available data. While exploring linguistic data, the students may form hypotheses and/or test them against the corpus data, or they may generalize or discover new

rules/knowledge from such resources. Accordingly, the metaphorical concepts typically combined with DDL are '*discovery learning*' and '*learners as researchers*'. The researcher-like methodology of DDL is viewed as an important way to engage learners in the world of language knowledge.

The concepts of DDL, discovery learning and learners as researchers are typically put into action through the use of language corpora and concordance-based methods. Leech (1997, p.3) mentions, "The student-centered paradigm of *discovery learning* – or what John has called *data-driven learning* – can scarcely be better exemplified than through the use of the computer corpus". In DDL, data are naturally occurring texts so both teachers and students do not necessarily know what will be found in the corpus. In practice, learners become researchers into language, forming hypotheses and testing them against the authentic data provided by the corpus and the teacher becomes a research organizer. Robinson (1991, p.3) concludes, "An open-ended and uncensored supply of language data encourages students to explore and discover rules for themselves with the guidance of the teachers".

In DDL, it is assumed that learning takes place either deductively or inductively when students are engaged in language analysis from a particular corpus. The concordance-based method demands students to identify examples in the corpus that match particular categories or types (deductive approach), or to come up with patterns or generalizations (inductive approach). In other words, in a deductive approach, the teacher firstly teaches rules which can be tested and confirmed by the data. Then students are assigned to find evidence from authentic language use in the corpus. The rules are set as hypotheses for students to test whether the evidence from corpora will support or deny the given rules. It is believed that students are unlikely to understand the rule statement until they have tested it against various examples. However, McDonough (1995, cited in Stevens, 1995) argued that giving a rule first imposes a rule formation rather than encouraging the student to make one up in his own terms. On the other hand, without rule teaching in an inductive approach, learners explore available data from corpora to generalize or induce rules/patterns. Supporting an inductive approach, Shaffer (1989, cited in Todd, 2001) describes the nature of induction as a two-stage process. Firstly, learners focus their attention on examples illustrating the target language point, and then they consciously generate rules or patterns from these examples. To investigate students' inductive ability, Todd

conducted a study with Thai postgraduate students. In his study, the students used self-selected concordances to carry out self-correction on their own writing. It was found that concordances could facilitate inductive learning because the students were generally able to induce valid patterns from their self-selected concordances and make valid self-corrections of their errors. However, despite much research on induction versus deduction, it is still unclear as to which one is more effective.

Currently, a shift in ELT moves towards a more learner-centered paradigm of '*discovery learning*', more pedagogical activities have required learners' involvement in order for them to make their own discoveries (Tomlinson, 2002) The role of the classroom and teaching materials is asserted to aid learners to make effective use of the resources in order to facilitate self-discovery. Similar views are also expressed in Todd (2001), Gavioli and Aston (2001), Thurstun and Candlin (1998), Fox (1998), Willis (1998), Cobb (1997a and b), Stevens (1995), and Tribble and Jones (1990). Therefore, it can be stated that '*discovery learning*' of the DDL approach conforms nicely with such currently popular ideas in ELT approaches, rather than being an activity in which knowledge is simply handed down from teacher to students.

#### **2.5.4. Compatibility to various approaches to vocabulary instruction**

Although the concordance-based method is closely associated with DDL, it is not necessarily restricted to only DDL. Its application can be used compatibly with all methodologies concerning explicit instruction (Gabrielatos, 2005). Equipped with a corpus and a concordancer, the concordance-based method is potentially applicable in vocabulary instruction. It is very promising in terms of arranging optimal conditions for enhancing vocabulary acquisition. Firstly, it can provide a unique resource of authentic and representative language specifically serving students' needs. Secondly, frequency wordlist can be quickly created to be used as one criterion in selecting target words. Thirdly, words can be easily contextualized since thousands of words in multiple contexts can be searched and obtained easily and quickly with a concordancer. With this method, the materials can be prepared much more conveniently and quickly in order to present target words in various authentic contexts with ample encounters of authentic language samples. This introduces learners to a large number of target words in a short time. Therefore, vocabulary

learning in multiple contexts can be facilitated and a sufficient number of word encounters can be achieved.

Apart from providing optimal conditions, the corpus-based method fits well with the current approaches of vocabulary instruction i.e. explicit, strategy and implicit instructions as well as teaching vocabulary through reading as discussed in 2.4. In *explicit instruction*, target words can be made salient with the display of KWIC concordances where the target words are presented in the center with the contexts on each side. Accordingly, students' attention is drawn directly to the words being studied and at the same time students can easily observe word behaviors in multiple contexts. Drawing students' attention to target words conforms to '*noticing hypothesis*' by consciously focusing on both forms and meaning of words. Based on *noticing hypothesis*, what learners notice in input is what becomes intake for learning (Schmidt, 1995, p.20). Moreover, being assigned to observe contexts of keywords, students can study various features of each of the words to be learned such as its grammatical functions, various meanings, collocations etc. Meeting a word in different contexts expands lexical knowledge of that word with its various features and repeated encounters consolidate quality or depth of word knowledge. This accumulation of knowledge strengthens knowledge of a particular word. In addition, with the concordance-based method, the recycling of word encounters is very easy. Words newly learned likely become contexts of the next target words if such words are words frequently occurring in particular types or topics of texts. When *teaching vocabulary through reading*, the concordancer allows for referring to source texts, which can be a good resource for reading. With *intensive reading*, concordance vocabulary exercises associated with reading can be developed both in paper-based and computer-based formats. '*Narrow reading*' can be facilitated to enable low proficiency learners to access authentic texts by reading a number of short passages on the same topic. Short texts on the same topics can be compiled and stored in a corpus in order to ensure the recurrence of particular words and patterns.

In *strategy instruction*, the concordance-based method is also helpful in making the clue words salient. For example, students may inductively study words used as discourse markers by searching words such as '*however*' and '*known as*', and then infer how these words hint the meaning of unknown words. This is to strengthen students' *inference skills* in guessing unknown words from context clues. When

students can infer or generalize rules by themselves, it seems that they are carrying out '*self-discovery learning*'. On the other hand, students may study deductively by initially studying the ways these clue words give hints to the meanings of other words, and then searching a corpus to find examples to verify whether it supports or denies the rules. This is like a researcher trying to test hypotheses as in the metaphor of '*learners as researchers*'.

In *incidental learning*, however, the concordance-based method does not play a large role in the part of classroom activities since in this approach means the students are assigned to read a lot of books. Instead, it may be helpful in grading texts according to the level of difficulty. In *teaching vocabulary through reading*, the corpus-based method is possibly applied in both bottom-up and top-down paradigms. With a bottom-up approach, it deals with the exercises of discrete elements of reading similar to the activities mentioned above in the explicit instruction. With a top-down approach, a word frequency list of the reading passage is a good starting point for discussion about the topic in order to practise overviewing or predicting skills on the text topic before reading. Vocabulary exercises may be inserted into the reading practice when feasible.

In the present study, the application of the concordance-based method is not restricted to any particular methodology. Therefore, the approach of DDL can be modified and expanded to incorporate other teaching techniques to increase the potential of the concordance-based method. In addition, the focus of the lesson can be made more flexible for facilitating students' direct access to the corpus (Gabrielatos, 2005). Accordingly, learning processes in dealing with the concordance output may vary from one situation to another, depending on the objectives of the study as well as the selected teaching techniques. With low-proficiency students, the present study adopted a deductive approach of DDL in explicitly teaching vocabulary through reading in concordance lines. This framework demands two basic skills of the students: skills in dealing with a computer concordancer to facilitate the observation of word behaviours in contexts and skills in dealing with concordances to learn vocabulary.

### **2.5.5 Previous application of the concordance-based method in ELT**

At its early stage in ELT, the concordance-based method was applied exclusively among developers of curricula, syllabuses and materials. Recently,

however, this method has increasingly applied directly in language classroom. It was not until Johns (1991) originated the approach of data-driven learning inherently associated with the concordance-based method that more empirical studies have been found. These previous studies are reviewed according to their types of work: development of syllabuses and materials, and classroom activities in order to overview the status of classroom concordancing in ELT as well as to find areas where further research is required.

### **2.5.5.1 Development of syllabuses and class materials**

The interest in the use of authentic materials in language pedagogy has enhanced the role of corpora in language instruction because corpora are enormous resources of real language use. Moreover, the demand for specific purpose language further increases the use of corpora in order to identify specific language in particular target situations. In addition, research based on corpus evidence indicates the need to exploit corpus information in developing curricula, syllabuses and class materials. Findings from such research reveal that the standard account of certain grammatical patterns in English does not conform to those that are actually used in natural language (Conrad, 2000; and Fox, 1998). The grammatical usage of ‘*some*’ and ‘*any*’ is one example. In traditional grammar books, ‘*some*’ is prescribed to be used in statements whereas ‘*any*’ is used in interrogative and negative sentences. However, it is found from corpus-based data that ‘*any*’ is much more frequently used in statements than the other patterns. Such discrepancy between prescribed grammatical usage and real language use has been increasingly found from corpus information. Previous studies (Fox, 1998; Willis, 1998; and Carter, Hughes and McCarthy, 1998) suggested that content selection and grading as well as materials development needs to be informed by some degree of corpus studies. It is argued that syllabuses and materials derived from concordance output in the corpus best serve pedagogical needs for guiding learners while still providing authentic and representative language.

However, the degree of text authenticity has been one controversy in the field of EFL instruction, according to Guariento and Morley (2001) and Tomlinson (2002), “One side argues that simplification and contrivance can facilitate learning; the other side argues that they can lead to faulty learning and that

they deny the learners opportunities for informal learning and the development of self-esteem”, (Tomlinson, 2003, p.5). On the one hand, it is considered that pedagogic simplification of real language use is necessary in order to protect learners from the apparent chaos of reality and to provide a sense of progress. In most cases, to avoid syntactic complexity, language input is constructed or composed rather than representing authentic contexts so that learners can focus their attention on the target language features (Bloor, 1998). On the other hand, the counter argument is that language inputs with composed texts over-protect learners and do not prepare them for real language use. Learners also need to be prepared for interaction in real situations. The lack of conformity makes English language in classes insufficient for preparing students to cope with their academic language in real situations. Students cannot be trained sufficiently if classroom English is learnt in one way but real language is used in another. In order for them to perform well in accomplishing academic tasks, language input must be similar to real language occurring in authentic texts.

At present, findings from large-scale corpora have been utilized to inform syllabuses and materials development. Word lists as discussed in 2.1, for example, are based on huge corpora which help define goals for vocabulary learning. In addition, Flowerdew (1993) demonstrates how to exploit a small corpus to draft a syllabus for a particular domain whereas Fuentes (2001) describes how the results of the contrastive study of lexical items in small specific corpora can become the basis for teaching/learning ESP at the tertiary level. More books are making use of corpus data reflecting actual language use rather than using non-authentic input. Stevens (1991a) develops vocabulary materials derived from relevant authentic texts in the corpus whereas Thurstan and Candlin (1998) used concordancing programs to develop academic materials for independent learning. Class exercises and activities in corpus-based materials are continuously designed. Typical exercises include vocabulary building, exploration of grammar and discourse features of texts. Specific description for designing classroom materials and exercises can be seen in the work of Tribble and Jones (1990), Fox (1998), Willis (1998), and Thurstan and Candlin (1998). In such works, concordancers exemplify how to highlight grammatical patterns, collocations and pragmatic aspects of lexical items. They are used as one form of text manipulation. The concordance output can be easily converted into



teaching materials, either by editing with a word processor or by old-fashioned scissors and paste methods. These studies provide useful frameworks for the development of syllabuses and class materials.

### 2.5.5.2 Classroom activities

Current teaching methods have emphasized the importance of aiding learners to make effective use of the resources in order to facilitate self-discovery. Therefore, learning activities in which learners can access and make use of corpora in language learning have become more popular, and the use of such hands-on corpus-based activities in classroom is encouraged. This type of application can be summarized in two main aspects: a soft version and a hard version (Gabrielatos, 2005). In a '*soft version*', learners do not have a direct contact to language corpora, only using the paper-based materials derived from corpora which are prepared by the teachers (Sriphicharn, 2002; Fuentes, 2001; Fox, 1998; Willis, 1998; Carter, Hughes and McCarthy, 1998; Thurstan and Candlin, 1998; Flowerdew, 1993; Stevens, 1991b; and Tribble and Jones, 1990). On the other hand, in a '*hard version*', learners conducted hands-on activities to utilize corpus information for their learning (Chan and Liou, 2005; Kaur and Hegelheimer, 2005; Hadley, 2002 and 2001; Cobb and Horst, 2001; Todd, 2001; Cobb, 1997 a and b; and Somogyi, 1996).

Earlier related works in classroom concordancing were concerned mostly with evaluating the concordance-based materials or programs; and/or giving practical frameworks, guidelines and suggestions for applying the concordance-based method in the classrooms, rather than providing empirical evidence. Owen (1997) and Fox (1998), for example, advocate the use of corpora as a reliable reference. They argue for both teachers and learners to directly consult evidence in a corpus, rather than relying only on what is grammatically prescribed. According to them, differences are found between language use in classrooms and the use in the real world. As being evident from the corpus, there are more varieties of real language use than grammatical usage in classroom. Thus, consulting evidence in a corpus is much more reliable.

An application of concordances to ELT classroom activities is proposed by Tribble and Jones (1990). They trace the history of concordances from the 13<sup>th</sup> century when concordancing was originally a paper-based method of

analyzing culturally valuable texts, and describe the features and the application of classroom concordancing with clear illustrations so that ELT teachers with no experience in this program can easily understand them. They also illustrate the utilization of concordance output in designing teaching materials as well as hands-on activities.

Specific descriptions of classroom activities can be seen in the works of Johns (1991), Fox (1998), and Willis (1998). In these works, concordancers are used to highlight grammatical patterns, collocations and pragmatic aspects of lexical items. They also serve as one form of text manipulation. All these researchers agree that learner training for concordancing application is necessary to prevent learners' confusion and prepare them to explore tasks. Some examples of training exercises are also provided. It is believed that once learners have become confident at using concordancers, they can develop their own research project. All their sample exercises provide good models for the practical application at a starting point.

Unlike computer concordancing, Willis (1998) suggested hand-concordancing, with concordances written on the blackboard. This type of concordance did not need the assistance of computer technology. The corpus derived from texts familiar to the learner from which linguistic patterns were selected by learners and used to compile concordances on the blackboard. She called it a '*pedagogic corpus*'. Similarly, Todd (2001) also used web-based texts for students to do hand-concordancing. The students '*self-selected*' the web-based texts to carry out hand-concordancing on their error words in report writing and then, tried to induce the meaning of these words for self-correcting these errors. It was found that the students could induce valid patterns in the carrying out of self-correction.

It was not until Johns (1991) originated the approach of data-driven learning inherently associated with the concordance-based method that more empirical studies have been carried out. Stevens (1991b) conducted a controlled experiment on students' offline concordancing. The students' task was to recall a known word to fill a gap in a text, either a gapped sentence or a set of gapped concordance lines for a single word. It was found that students could retrieve a word from memory more successfully when cued by the concordance lines. John (2001) used a parallel corpus with a concordancer. The study was aimed at determining how

students dealt with the parallel corpus and what conclusions they come to when comparing the two languages, and in particular when investigating lexical items.

One distinctive series of studies was conducted by Cobb (1999a and b; and 1997a and b) as well as Cobb and Horst (2001). He developed a computer concordance-based tutoring program called *PET 200* in his doctoral study and the later version of *PET 2000*. These programs were tested with undergraduate students at Sultan Qaboos University in Oman to assess their learning effects on students' definitional and transferable knowledge. Hands-on activities were used for students to access the given corpus in order to accomplish the assigned tasks independently. Target wordlists were set for students to self-select words they would like to learn. Findings from Cobb's series of studies consistently showed that the concordance-based method could increase students' vocabulary knowledge significantly in a short time, especially transferable knowledge. In other words, hands-on concordancing helped them acquire more transferable knowledge.

Concordances were also successfully applied in academic writing (Todd, 2001; Tompson and Tribble, 2001; and Webber, 2001). Concordancers were used as references for students to discover their own weaknesses or errors in writing and then improve them. Webber (2001) advocated a concordance and genre-based approach to academic essay writing for non-native students. Students were required to identify some structural characteristics of a legal essay. Then they used concordances to explore possible correlations between the generic structures and particular lexical items. Next they were asked to rewrite an essay. Thomson and Tribble (2001) focused on citations from a corpus of doctoral theses. They introduced a number of class activities in which students conducted their own analyses of citation practices in small corpora, to develop genre awareness.

In two case studies of Hadley (2001 and 2002), data-driven learning was introduced to Japanese students using the paper-based *Concordance Sampler 2* as class materials. The students were exposed to a large amount of the pre-selected concordance materials to identify regularities in data for application to writing tasks. In both studies, students' attitudes were positive towards DDL, although the concordance materials were rated as '*difficult*' due to a large amount of data and the difficulty level of the materials.

In more recent studies, Chan and Liou (2005) and Kaur and Hegelheimer (2005) used web-based concordancers, *TOTALrecall* and *Compleat Lexical Tutor*, to investigate the learning effects on collocation learning and transfer of word knowledge to writing tasks respectively. Significant gains from the web-based concordancing were found in Chan and Liou's study but students' retention of knowledge was weaker although the residual effect was significantly high. In Kaur and Hegelheimer's study, students' performances in vocabulary tasks of the experimental group were not significantly different from those of the control group. However, they outperformed the control group with significantly different transfer of vocabulary knowledge applied to the writing tasks.

Concordance-based activities were also found in Thai educational contexts in two classroom-based studies of Sripicharn (2002) and Todd (2001). Both studies were conducted with university students. In Sripicharn's study, the use of DDL or classroom concordancing was evaluated in three aspects i.e. its learning effects, learners' attitudes and learners' performance during the use of classroom concordancing. Concordance materials were used in the experimental group about 10-15 minutes near the end of each lesson whereas the control group used non-concordance materials. It was found that students were able to make useful generalizations and adopt DDL in dealing with the concordance data. Although there was no marked difference in the learning effect between concordance and non-concordance methods, students' attitudes towards classroom concordancing tasks were positive. In a similar vein, in Todd's study, the method of self-selected hand-concordances was used for students' self-correction. Errors in students' written assignments were highlighted for students to correct themselves by comparing their use with that occurring in the concordances they selected from the Internet. It was found that the students were able to do self-correction by using self-selected concordances and their attitudes towards the method were positive.

## **2.6 The Present Study**

Based on the literature review in this Chapter, there does not exist a great amount of empirical evidence concerning classroom concordancing due to its recent

introduction to ELT. Therefore, there is plenty of research space in this area. The present study is distinctive from other previous studies at least in five aspects. Firstly, it has a different aspect of classroom application. In previous works, the method was mostly applied as a referential tool (Chan and Liou, 2005), as parts of the courses or supplementary to other teaching methods (Sriphicharn, 2002) for correcting errors in report writing (Todd, 2001) and/or as self-access or tutorials (Cobb, 1999a and b; and Cobb, 1997a and b). Only Kaur and Hegelheimer's (2005) study was implemented in regular class time with the focus on using concordances in the transfer of academic word knowledge to writing tasks. In contrast to these studies, the concordance-based method in the present study is applied as the main method fully integrated with the whole regular course in one academic semester.

Secondly, concordancing facilities in the present study are also different. Previous studies with hands-on concordances were mostly conducted with experimental or specifically designed concordancing programs: Tom Cobb's web-based Compleat Lexial Tutor in Kaur and Hegelheimer (2005) to learn verb-noun collocations; PET-2000 in Cobb (1999a and b) and PET-200 in Cobb (1997a and b) to learn words for Cambridge Preliminary English Test (PET); and TOTALrecall of Liou, Chan and Yeh et al. in Chan and Liou (2005) to learn academic words for writing. In contrast, the present study uses a simple freeware program, WCONCORD, with a small specialized corpus specifically compiled from engineering academic texts to select target words as well as to design learning materials and activities. This does not only make the concordancing application more practical in normal classroom practices, but it also makes the concordance-based lessons in the present study distinctive because target words can be contextualized within academic texts in engineering. Therefore, the students can learn these words in their familiar contexts, which is likely to encourage them to learn. Very few works have been done in developing concordance class materials for technical or engineering students, even in traditional paper-based textbooks.

Thirdly, the method of word selection and the designs of concordance-based materials and activities are based on various techniques from current teaching in order to enhance the effectiveness of the concordance-based method as well as to make its application conform to normal classroom practices. Although the method is related to an approach of Johns' (1991) '*data-driven learning*', Gabrielatos (2005) points out

that its application is not necessarily restricted only to any single teaching methodology since it can be compatible with all methodologies that accept explicit learning/teaching. In addition, learning materials and activities are designed by using the contexts of engineering to serve the needs of the students in the study although the focus is on learning academic words in general. So far, these designs have not been found in any published papers.

Fourthly, the empirical data derived from the present study are distinctive from those of other previous studies in terms of learning outcomes. The study is aimed at dealing with three levels of lexical knowledge: definitional knowledge, transferable knowledge and retention rates. Compared to other vocabulary research, although Cobb's studies (1997a and b), also dealt with learning gains in definitional and transferable knowledge, the retention rates of such knowledge remained unknown since most of his studies were mostly conducted with one group of students, using only the pretest and the immediate posttest. In the present study, however, retention rates of both knowledge types after a month of the study are also estimated with a delayed posttest in comparison with those of the conventional teaching method. In addition, the focus of the present study is on academic vocabulary for academic reading, which included high frequency words in the engineering corpus whereas the focus of Cobb's studies was on vocabulary for PET. Regarding another study (Kaur and Hegelheimer, 2005) dealing with the transfer of academic knowledge, the focus of the studies are not identical since the present study focuses on transferring vocabulary knowledge to new academic reading contexts whereas the other study focused on transferring academic vocabulary to writing tasks.

The fifth aspect is that this present study explores the first classroom-based research, using hands-on concordancing in Thai educational contexts. So far, two classroom-based studies (Sriphicharn, 2002; and Todd, 2001) in Thailand have been published. Both studies, however, used paper-based concordances and hand-concordancing to supplement other teaching methods, not as the main method. Therefore, the present study provides original empirical data of using hands-on concordances as the main method of teaching vocabulary for academic reading in Thai classroom contexts.

Finally, many previous classroom-based studies were conducted with one group of students (Chan and Liou, 2005; Hadley, 2001 and 2002; Todd, 2001; Cobb

and Horst, 2001; and Cobb, 1999a; Cobb 1999a and b), regardless of a control group. Therefore, the applications of the concordance-based method in these studies were not investigated in comparison with those of other teaching methods. Only a few studies have aimed at making such comparisons – the applications of a concordancer together with an online dictionary and only the online dictionary (Chan and Liou, 2005), the paper-based concordance and non-concordance materials (Sriphicharn, 2002), and a concordancing program and a wordlist with a dictionary (Cobb, 1999b). To provide empirical evidence in this gap, the present study is aimed at comparing the learning effects of the concordance-based method and the conventional teaching methods on learning vocabulary through reading, using the contexts in reading concordance lines in the former method and reading sentences or passages in the latter method.

To summarize, the present study attempts to bridge the gaps in this area of research. The effects of a hands-on concordance-based method are used as the main method in comparison with the conventional teaching method on vocabulary learning in the whole regular course. The application of a simple concordancer and a small specialized corpus can provide a practical framework for most EFL academic situations, especially with engineering students. The findings of the study originally provide details about the effects of using the hands-on concordancing method in Thai classroom contexts. Accordingly, these findings of the study contribute to the area of teaching English for Academic Purposes (EAP) in providing useful implication as well as empirical evidence in the areas where research is lacking.

## **2.7 Summary**

In this chapter, academic vocabulary is described in terms of word classification, lexical thresholds for academic reading and the assessment of vocabulary size. These matters are very useful in pedagogy for identifying and selecting target words suitable for the objectives of the study as well as in designing the instrument properly for assessing knowledge of the target words. Previous frequency-based studies have classified words into four types: high frequency words, academic words, technical words and low frequency words. Apart from word

classification, frequency-based lexical thresholds have been established for academic reading, reading comprehension and vocabulary size. Currently, the established wordlists of the GSL and the UWL/AWL, which consist of around 3,000 word families, are widely acceptable as a lexical threshold for academic reading in all disciplines. As a result, measures assessing such vocabulary have been developed both in the forms of receptive and productive versions.

The literature on vocabulary acquisition and retention provides insights into how to enhance students' vocabulary learning and retention. In this topic, two vocabulary knowledge types: definitional and transferable knowledge are described: the former is knowledge of meaning whereas the latter is concerned with the ability to transfer lexical knowledge to new contexts. Regarding the nature of vocabulary acquisition and retention, learning particular words is incremental. To learn a word well is to learn various aspects of word. Thus, meeting a word only once is not sufficient for that word to be learnt properly. At the first encounter, definitional knowledge usually takes place since it is at a superficial level of knowledge concerned only with recognizing word forms and memorizing their meaning. For knowledge transfer, however, that word must be accumulatively learnt. Quality of lexical knowledge takes place and is strengthened through meeting the to-be-learnt words in a variety of natural contexts several times before such knowledge can be transferred to new contexts. Therefore, learning words in multiple contexts with a sufficient number of word encounters seems to be essential for vocabulary acquisition.

In ESL/EFL contexts, the status of vocabulary has changed over time before it becomes prominent at present. The vast improvement of corpora greatly contributes to vocabulary studies. Findings from corpus-based studies have reoriented perception on nature of language and vocabulary, and have made the status of vocabulary more distinct. Three teaching approaches have been developed with the increasing attention to vocabulary instruction, ranging from no instruction to formal instruction. The advantages of these approaches can be integrated into '*teaching vocabulary through reading*' and such integration seems to be the best approach to obtain optimum vocabulary learning.

With the rapid advancement of computer corpora, the corpus-based method was introduced to language instruction with a new promising direction of pedagogical



practice. Frequency-based lexical thresholds and reliable wordlists are established.

Words are selected for designing syllabuses, materials and classroom activities in a much more systematic way. Despite being new, the method conforms well to principles and methods of current teaching/learning approaches. It is useful not only in arranging optimal conditions for vocabulary acquisition, but also in arranging classroom activities. However, empirical evidence of its application in classroom is not abundant and models for such application with various levels of students are still needed. The present study is an attempt to apply the concordance-based lessons with undergraduate students with low proficiency. It is aimed at providing a practical framework as well as empirical evidence on the learning effects of the method, which currently are very rare.