VocabAnalyzer: A Referred Word List Analyzing Tool with Keyword, Concordancing and N-gram Functions*

Siaw-Fong Chung[a], F.Y. August Chao[b], and Yi-Chen Hsieh[a]

[a]Department of English, National Chengchi University,
No. 64, ZhiNan Road Section 2, Wenshan District, Taipei City 11605, Taiwan
{sfchung, 96551016}@nccu.edu.tw

[b]Department of Management Information Systems, National Chengchi University,
No. 64, ZhiNan Road Section 2, Wenshan District, Taipei City 11605, Taiwan
fychao.tw@gmail.com

Abstract. This paper introduces the newly created *VocabAnalyzer* which is equipped with keyword, concordancing and n-gram functions. The *VocabAnalyzer* also allows the comparison of the inputted text against Jeng et al. (2002) vocabulary word list. Two case studies will be discussed in this paper. The first study compares two versions of the English Bible and the second study compares word list created by abstracts written by graduate students of various English departments in Taiwan. Both study shows that the *VocabAnalyzer* is beneficial for applied linguistics studies as well as for teaching and learning of English. Analyses of student writing can be carried out based on the results from this tool.

Keywords: *VocabAnalyzer*, word list, keyword, concordancer, vocabulary

1 Introduction

With the increasing expansion of corpus linguistics and its applications to fields such as theoretical linguistics, applied linguistics and language teaching, concordancers are now in greater demand, especially by researchers who do not write programming scripts themselves. Wordsmith (Scott,1999), AntConc (Anthony, 2004) and the Concordancer Software (Nguyen and Munson, 2003) are some of the most frequently used concordancers because these programs allow plain texts to be processed and displayed as KWIC or analyzed according to collocations. Therefore, one can easily create a collection of texts to be processed by these tools. Other concordancers which were designed for specific purpose are such as Concgram (Greaves, 2005) (for phraseology), ParaConc (Barlow, 2002) (for parallel corpora), Collocates (Barlow, 2004) (for collocation extraction) and Xiara (Burnard, 2004) (for reading XML files). All these

* The research reported in this paper was supported by a grant from the National Science Council, Taiwan (NSC 97-2410-H-004-001-) to the first author. The authors would like to thank Chun-Hung Chen and other members of the Corpus-based Research Group, NCCU, who have helped in different ways under this project. Dr. Zao-Ming Gao (unpublished) has also worked on an interface comparing words in an inputted text against Jeng et al.'s (2002) word list (http://140.112.185.57/~mahanaim/cgi-bin/readability1.0.pl). His interface focuses on the calculation of readability tests and presentation of statistics about an inputted text. The *VocabAnalyzer*, however, goes for a user-based interface with incorporation of keyword, n-grams, and concordancer functions accompanied with some basic statistics about the inputted text. Readability tests are not performed by the *VocabAnalyzer*. The first author would like to acknowledge and thank Dr. Gao for the insightful comments and recommendation about Jeng et al.'s word list at the early stage of this project, which becomes one of the reasons behind the motivation of this work.

concordancers allow uploads of texts for specific analyses. Most of these tools come with a concordancer function too.

In this study, we use a newly created online interface for comparisons of vocabulary in different texts. Previously, there are also vocabulary tools such as AntWordProfiler (Heatley, Nation and Coxhead, 2002) and Range (Nation and Heatley, 2003), both of which are based on Paul Nation's Word List. Not many of these have combined a vocabulary analyzer with a concordancing function. The *VocabAnalyzer* (http://metcon.corpusresearch.nccu.edu.tw/vocab) is an open source (to date) with several functions: (a) a display against a six-leveled vocabulary list built in Taiwan by Jeng et al. [鄭恆雄等] (2002) — a vocabulary list created based on teaching materials used in Taiwan; (b) some basic statistics about the number of sentences as well as type and token counts; (c) a keyword function enlisting all words from an inputted text; (d) a bi-gram and tri-gram system for an inputted text; and (e) a concordancer function with a query system based on the words in an inputted texts. These functions are themselves no new in corpus linguistic research but when they are put together with a purpose of applied linguistic research, the *VocabAnalyzer* becomes a benefit to both corpora and vocabulary researchers. In addition, since context is one of the significant factors in learning, Jeng et al.'s word list which was created exclusively for English learners in Taiwan, is, therefore, advantaged because of its direct application to teaching and learning in the Taiwan EFL contexts. Even though Nation's word list is highly acclaimed, the *VocabAnalyzer* will become handy for scholars and students in Taiwan especially those working on applied (corpus) linguistics and teaching and learning. It will also be a tool for comparisons of word list from Taiwan and that of Paul Nation's.

2 Leveled Reference Vocabulary List

Jeng et al. (2002) created the six-leveled English word list from English EFL textbooks. The following is an excerpt from the abstract of the word list (2002: i-ii; available from http://www.ceec.edu.tw/Research/paper _doc/ce37/ce37.htm).

> This English reference word list contains 6,480 words in American English, which were selected by referring to three kinds of materials: (1) nine sets of junior and senior high school English textbooks published in Taiwan; (2) five sets of American elementary school English readers; (3) twenty-one English word lists compiled in the U.S.A., the U.K., Canada, Japan, Mainland China and Taiwan.

The following Figure 1 provides a snapshot of the words in level one, the most basic level of the six levels. Each level has 1,080 words in total, excluding the varied forms such as 'April/Apr.' and 'a/an,' shown in Figure 1. The bracketed number indicates a varied form (usually noun and verb) of the same word.

LEVEL 1 (1,080 words)

a/an	ant	beach
able	any	bear (1)
about	anything	beat
above	ape	beautiful
according to	appear	beauty
across	apple	because
act	April/Apr.	become

Figure 1: Snapshot of the Reference Word List

One of the functions of the *VocabAnalyzer* is to show matches of words in an inputted text against the word list. However, since the original list shown in Figure 1 above was not created in a machine-readable format, steps were taken to standardize all six lists. The detailed steps to produce the *VocabAnalyzer* are documented below.

3 Steps to Produce the *VocabAnalyzer*

Each section below illustrates the creation of the functions of the *VocabAnalyzer*.

3.1 Creating a Program-Readable List

The following steps were taken to produce program-readable word list. First, we converted the original six-leveled word list to a program-readable word list. Some of the words in the original word list contain slashes and bracketed number or word. The converted program-readable list enlists words in vertical format (see Table 1).

Note also that all converted words are in lower case to facilitate the match between an inputted text and the word list. In converting all words to lowercase, new line expressions such as '\n\r', '\n', and '\r', and all non-characters especially the bracketed number were also removed by using the regular expression '/[^\w\s]+|\d+/.'

Table 1: Creation Program-Readable Word List

Example 1:

Original Word List	`MRT/mass rapid transit/subway/underground/metro`
Converted Word List	mrt mass rapid transit subway underground metro

Example 2:

Original Word List	`measure (1) (ment)`
Converted Word List	measure measurement

This step created more entries than the original word list has. For instance, 'measure' and 'measurement' were counted separately for 'measure (1) (ment)' of example 2 in Table 1. Thus, the *VocabAnalyzer* calculates a match between words in an inputted text with the word list based on the accumulated frequencies for the separated entries. For example, for 'measure (1) (ment),' its occurrences in the inputted text will be the sum of frequencies from 'measure' and 'measurement.' The process of matching both texts is described below.

3.2 Matching an Inputted Text against the Word List

When an inputted text is entered onto the interface, it will first be tokenized. Tokenization follows the logics of the NLTL toolkit (Bird and Loper, 2004) (http://www.nltk.org/) and the tokenized words will be lemmatized according to the Wordsmith e-lemma word dictionary (cf. Scott, 1999) (e.g., 'go' is the lemma for 'goes,' 'going,' 'went,' and 'gone'). After lemmatizing the tokens, the matching of the tokens with the word list will be carried out. The program then calculates word frequency and classifies words in to six levels. An additional unfound level is also created if no-match is produced. An output from *VocabAnalyzer* can be seen in Figure 1 below.

A pie chart on the left illustrates the distributional overview of the six levels plus the unfound level. This column also provides a simple statistics about the texts in which tokens, types, type-token ratios, number of sentences and average sentence length are displayed. On the right is a word distribution table which presents classified words according to the six levels (from more basic to more difficult) as well as the unfound words at the bottom. Note that the unfound words show the non-lemmatized words from the original text. The words shown in the six levels are words according to the forms shown in the original Jeng et al.'s word list such that exemplified in Figure 1. The bracketed numbers indicate the sum of the varied forms

previously mentioned in Figure 1. The same form of word for verb and noun (e.g., 'shop') will be considered as one form as matching of parts-of-speech was not programmed in the *VocabAnalyzer*.

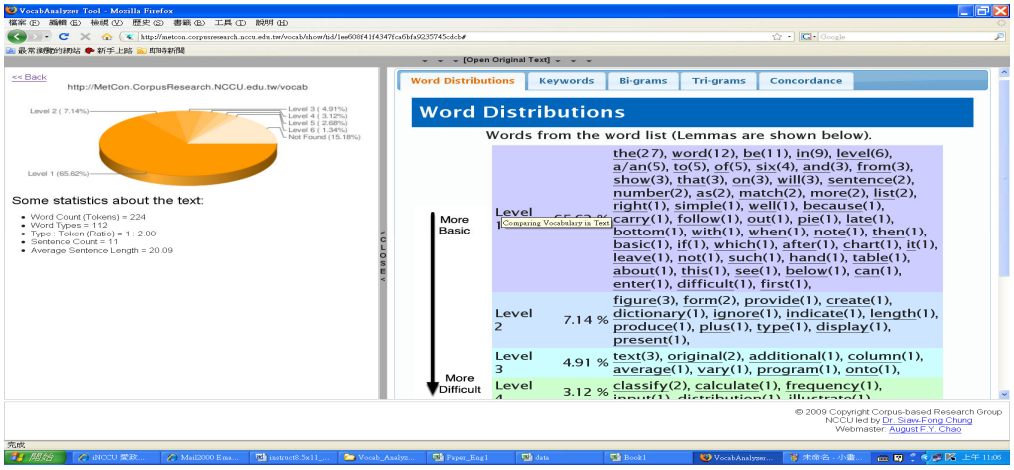

Figure 2: Snapshot of the *VocabAnalyzer*

All words listed on the right of Figure 2 can be linked to the concordancer function. This function is illustrated in sections 3.3 and 3.4 below.

3.3 Creating the Keyword Function

The Keyword function in the *VocabAnalyzer* enlists and summarizes all the words that are found in the inputted text (Figure 3a). The order of listing can be changed by clicking on the arrows (⟠) on the interface in Figure 3a. The clicking of any of the keywords will lead to its display of concordancer lines, as shown on the Figure 3b.

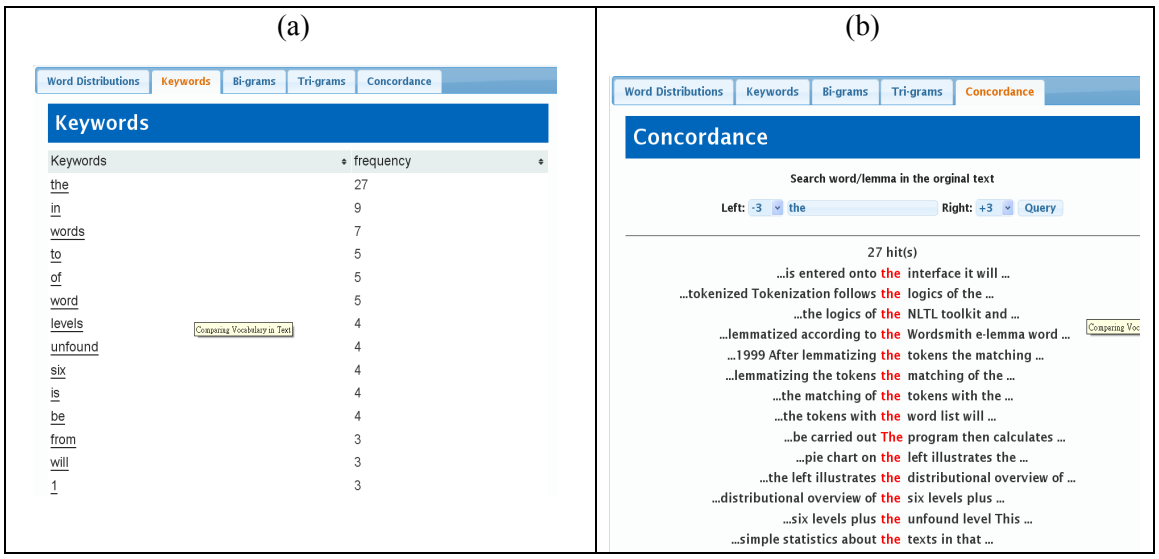

Figure 3: Snapshots of the Keyword and Concordancer Functions of the *VocabAnalyzer*

The creating of the Concordancer function is documented below.

3.4 Creating the Concordancer Function

The concordance function of the *VocabAnalyzer* can be linked through three paths – by clicking a word in the leveled word list (Figure 2), by clicking a keyword (Figure 3a) and by searching

for a keyword itself on the concordancer tab (Figure 3b). Since only the leveled word list displays words according to the forms in Jeng et al.'s word list, its path to the concordancer will need a mapping to the inputted text (so that the display on the concordancer uses words from the inputted text). After this mapping, all three paths follow the same procedures in displaying the concordance lines. The Wordsmith e-lemma word dictionary (Scott, 1999) is again used in the concordancer function so that both word and lemma can be searched in the concordancer function. Users can also select the desired windows size to adjust the display pattern.

3.5 Creating the Bi-gram and Tri-gram Functions

In the n-gram analyzers, the non-character strings of single quotation (') marks and dashes (-) were preserved. N-grams up till tri-grams are provided by the *VocabAnalyzer*. The bi-grams and n-grams were generated based on the NLTK toolkit. Examples of snapshots are given in Figure 4.

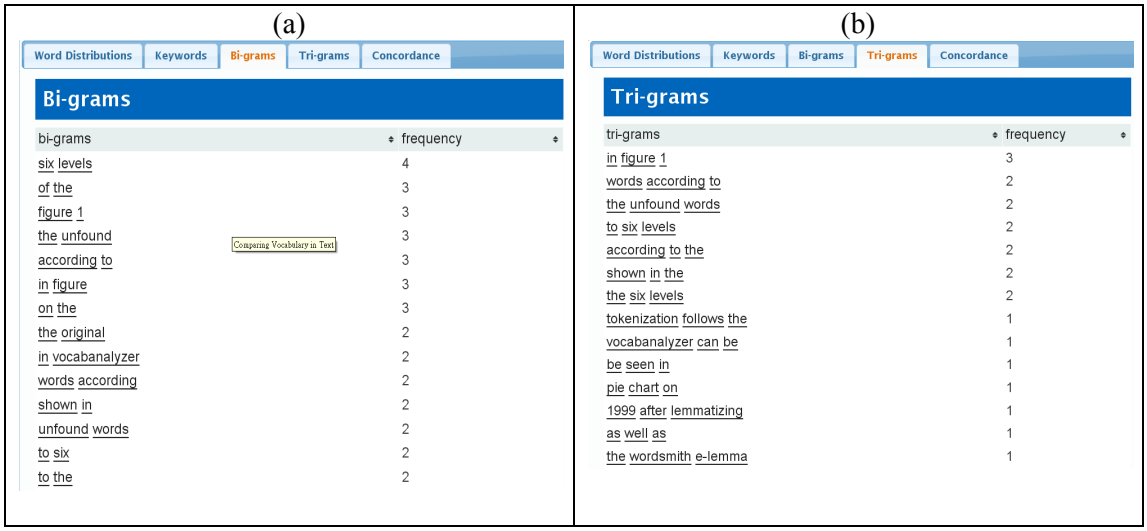

Figure 4: Snapshots of the Bi-grams and Tri-grams Functions of the *VocabAnalyzer*

The following section will discuss two case studies using the *VocabAnalyzer*.

4 Studies Using the *VocabAnalyzer*

This section will provide two case studies which can be carried out based on the *VocabAnalyzer*. The first study is an analysis of two versions of Bible (i.e. King James Version and Basic English Version), conducted in order to illustrate how the *VocabAnalyzer* could be adopted to provide pedagogical insights. Since different versions of Bible abound in the markets, the vocabulary items used may vary across versions, providing useful information in research and in language teaching. The second study is an analysis of advanced English learners' writing of abstracts. Such analysis informs teachers of their learners' command of the target English.

4.1 Comparisons of Two Versions of Holy Bible

The Holy Bible is the most common book to share religion knowledge with all ages of people, so the words should not be too difficult to understand. We analyzed chapter one of the book of Genesis in the Bible Corpus from University of Maryland Parallel Corpus (can be accessed through http://www.umiacs.umd.edu/~resnik/parallel/bible.html or through using the Almega BibleTools Library Version 3 (2009)). We inputted the bible chapter from the book of Genesis into the *VocabAnalyzer*, the output result showed that most words in the bible are located at level one (73.62%). Like most English texts, the bible also utilizes more words from this basic level.

The statistics information provided by the *VocabAnalyzer* in Table 2 show that the King James Version has fewer tokens but more word types in comparison with the Basic English Version.

Table 2: Statistics about the Holy Bible

King James Version 1:1-31	Bible in Basic English 1:1-31
Word Count (Tokens) = 797 Word Types = 135 Type : Token (Ratio) = 1 : 5.90 Sentence Count = 33 Average Sentence Length = 24.15	Word Count (Tokens) = 814 Word Types = 127 Type : Token (Ratio) = 1 : 6.41 Sentence Count = 30 Average Sentence Length = 27.13

The King James version contains 797 words whereas the Basic English Version comprises 814 words in total. With respect to the word type, the King James Version shows to have 135 counts whereas Basic English Version only has 127.

Information about the distribution of sentences could also be found. There are 33 sentences in the King James version and its average sentence length is 24.15 words in the King James Version. In contrast, the Basic English version contains 30 sentences and the average sentence length is 27.13 words. Therefore, the number of sentences in the King James Version is higher than that in the Basic English Version. However, the sentence is longer in Basic English Version. All the above information provides a picture of the distinction of the two versions of Bible.

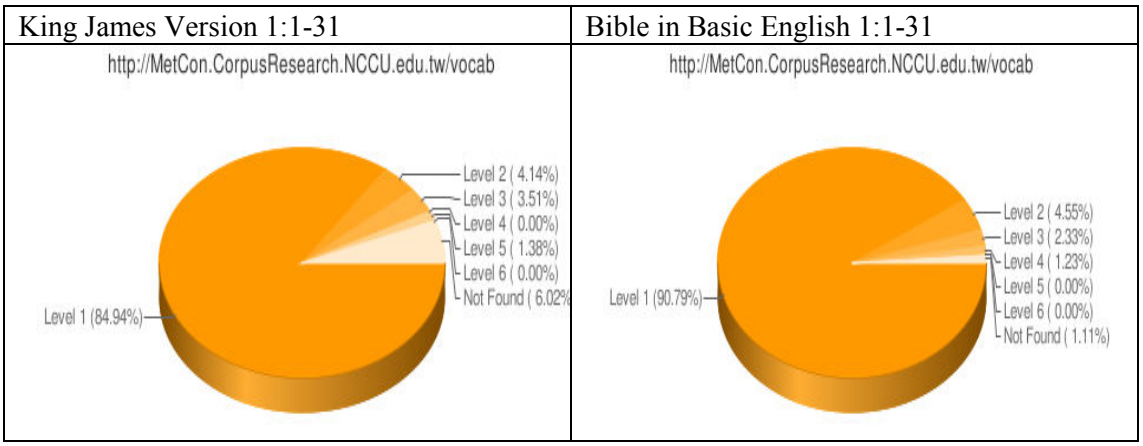

Figure 5. Comparison of the Vocabulary of Two versions of Bible

From Figure 5, we found that most of the vocabulary items appeared in the two versions are within the first three levels (92.59% for King James Version whereas 97.67% for Basic English Version). Possible reason may be that the Bible aims to be read by ordinary people and that most of the vocabulary items are not difficult. In addition, 92.59% of vocabulary items fall below level 4 in the King James Version whereas 98.9% fall below level 4 in the Basic English Version. We could infer that the vocabulary items introduced in the King James Version is harder than those in the Basic English version, a fact also known by many but we have proved that with the *VocabAnayzer*.

With regard to teaching, choosing the textbooks that suit students' needs is one of language teachers' duties. On top of the teaching intuition and experience, teachers could use the *VocabAnalyzer* as a vocabulary reference for textual analysis so as to have a better idea which textbooks suit which types of students. For instance, the above results from analyzing two versions of the book of Genesis (chapter one) tell us the difficulty level of lexical items used in two different texts. Based on the similar methodology, teachers can also choose their target

textbook based on such information. After choosing the textbook, teachers may consider which vocabulary items or phrases should be taught and what the presentation order of the chosen vocabulary or collocations is. Then, language teachers could resort to the contrast and comparison of the results of the bi-gram and tri-gram analyses. For instance, such results from the two versions may provide insightful information of which vocabulary items or phrases to teach and how to arrange them as the frequency data is one criterion used to make this decision. For example, teacher may choose to teach those highly-frequency words first. Furthermore, vocabulary and collocations included in both versions, such as 'earth', 'rule over' and 'every living thing', may be introduced first. In contrast, collocations or phrases, such as 'multiply', 'dominate over' and 'every living creature,' which could only be found in the more difficult version, can be introduced later when learners are more proficient.

This case study has provided an example from the teaching and learning perspective. In the second study, discussed below, an analysis of academic dissertation abstracts will be presented.

4.2 Analyzing Academic Dissertation Abstracts

Learners' production data may provide insightful information of how they command the target language. We analyzed 2,163 ETDS (Electronic Thesis Dissertations Service) abstracts of academic papers published by English majors in Taiwan from year 2003 to year 2008 in order to gain a snapshot of how learners used academic vocabulary items. Since all the abstracts are from Taiwan, Jeng et al.'s word list, which was specially-created for learners in Taiwan, may yield convincing results. The percentages of vocabulary according to different levels are presented in Table 3 below.

Table 3: Percentage of vocabulary of the 2,163 Academic Abstracts

Vocabulary Levels	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Not found	Total
Percentages of Coverage	17.84%	9.28%	6.79%	8.35%	3.64%	4.81%	49.30%	100%

After receiving training in the English departments, English majors are expected to possess a good command of lexical items of different difficulty levels. Result from Table 3 confirms this hypothesis. English majors could utilize both basic (33.91% is located from level one to level three) and advanced (16.98% is situated from level 4 to level 6) vocabulary when writing academic papers. The *VocabAnalyzer* also provides the figures of word counts. There are 111,283 tokens in total. In average, each abstract contains 51 tokens. Since the percentage of unfound words is high (49.30%), we then returned to the original texts and found that most of the unfound words are terminology. One example is the occurrences of abbreviations (e.g. *wbcp* which stands for Web-Based Career Project). Therefore, the unfound words from the *VocabAnalyzer* will also provide useful information about textual features. Such observation also offers language teachers where to emphasize by examining their learner's production. A refinement of this study, however, is needed so that comparisons between English-majors and non-English majors can be conducted. These two case studies (though the second one was briefly discussed) can be seen as examples regarding how the *VocabAnalyzer* will become a useful tool in research and in pedagogy.

5 Conclusion

As presented in this paper, the *VocabAnalyzer* allows comparisons of word lists and it also displays the keywords, n-grams and concordancing lines based on any inputted text. As a tool designed for the use of non-programmers, the tool is user-friendly in that its interface is intuitive. For research on testing and learning, the comparisons against Jeng et al.'s word list

will be a way to look into the types of vocabulary used. Chung and Wu (2009), for example, have examined the LTTC (Language Teaching and Training Center) Learner Corpus against the same word list too. With the applications provided by this, too, further comparisons of collocations are also made possible for their future work. For research on learner's writing, the tool allows automatic categorization of the learners' work in that correlation can later be performed between vocabulary levels and the scores given by a rater. For scholars not in the immediate Taiwan contexts who are not closely familiar with the word list, the *VocabAnalyzer* can also serve as a concordancer by itself. The tool, which is an open resource, thus, will become useful for linguists who intend to process their own corpus. Nevertheless, the *VocabAnalyzer* does not come with a part-of-speech tagging. In the future, this feature is hoped to be added to the tool.

References

[Jeng et al.] 鄭恆雄，張郇慧，程玉秀，顧英秀。2002。《大學入學考試中心高中英文參考詞彙表》。台北：財團法人大學入學考試中心基金會。

2009. Almega BibleTools Library Version 3 (Software). Almega System Analysts Limited. Available from http://www.almega.com.hk/main.asp

Anthony, L. 2005. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning,* pp.7- 13.

Barlow, M. 2002. ParaConc: Concordance software for multilingual parallel corpora. *Language Resources for Translation Work and Research, LREC 2002*, pp 20–24.

Barlow, M. 2004. *Collocate 1.0: Locating Collocations and Terminology*. Houston. TX: Athelstan.

Bird, S. and E. Loper. 2004. NLTK: The Natural Language Toolkit. *Proceedings of the ACL Demonstration Session,* pp. 214-217, Barcelona, Association for Computational Linguistics.

Burnard, L. 2004. BNC-Baby and Xaira. In TALC 2004: In Proceedings of the Sixth Teaching and Language Corpora conference, Granada, pp. 84-85.

Chung, S.-F. and C.-Y. Wu. 2009. Effects of Topic Familiarity on Writing Performance: A Study based on GEPT Intermediate Test Materials. Presented at *the 2009 LTTC International Conference on English Language Teaching and Testing*. National Taiwan University, Taiwan. March 6-7. [Full paper appeared in conference CD Rom].

Gao, Z.-M. Unpublished. *English Text Analysis* [Online Interface]. Available at http://140.112.185.57/~mahanaim/cgi-bin/readability1.0.pl

Greaves, C. 2005. Introduction to ConcGram. Presented at *Tuscan Word Centre International Workshop.* Certosa di Pontignano, Tuscany, Italy.

Heatley, A., I.S.P. Nation and A. Coxhead. 2002. *RANGE and FREQUENCY programs* [Software]. Available from http://www.vuw.ac.nz/lals/staff/Paul_Nation or as AntWordProfiler (Version 1.103, Windows and Macintosh) from http://www.antlab.sci.waseda.ac.jp/ software.html

Nation, I.S.P. and A. Heatley. 2003. *Range*. Victoria University of Wellington School of Linguistics and Applied Language Studies. Accessed July 21, 2004.

Nguyen, N. T. and E. Munson. 2003. The Software Concordance: A new Software Document Management Environment. In *Proceedings of 21st annual international conference on Documentation*, ACM Special Interest Group for Design of Communications SigDoc'03, San Francisco, California, USA, pp. 198-205.

Scott, M. 1999. *WordSmith Tools version 3.0*. Oxford University Press, Oxford, UK.