Developing a Profiling Tool for English Language Texts on Life Sciences: Compilation of a Bioscience Corpus and Its Application

Akiko HAGIWARA*, Mao NAITO**

Abstract

The present study explores the use of high frequency word lists to profile authentic reading materials intended to be used in the college-level English language program for bioscience majors. In order to accommodate the needs of students majoring in biomedical sciences, which require considerable proficiency of English with profound knowledge of scientific vocabulary specifically used in the field, using content-based reading materials in class is a good solution. It enhances students' vocabulary size and further helps them acquire state-of-the-art information in the fast moving field of science. Profiling English texts using a vocabulary list containing high frequency vocabulary in biomedical literature can facilitate the process of text selection in such content-based English language programs. Based on several types of corpora, a word list (*the LS Wordlist*) comprised of several sub-lists reflecting various types of vocabulary was generated. Three small corpora and nine texts were profiled using this word list. The output clearly indicates the characteristics of each corpus or text, showing how easy or technical the text or the corpus is for the potential learners. This study shows that by combining different types of corpora, it is possible to create a useful word list for profiling scientific texts.

Background

In English for specific purposes (ESP) programs, choosing appropriate reading materials for teaching is extremely important. One obvious reason is that a specific field has its preferred writing styles (Dudley-Evans and St. John 1998) and makes use of technical and sub-technical vocabulary specifically used in the field (Fraser 2001). Although it is ideal for a language instructor to choose such reading materials that contain appropriate proportions of these types of vocabulary and that meet the proficiency-level of the students at the same time, it is a very difficult task for several reasons. One is because at present there isn't a simple way to categorize what words are technical, sub-technical, academic, and so forth, so it is virtually impossible to classify any texts based on their vocabulary use. Another reason is that teachers often do not know how much vocabulary their students have already acquired, which makes it difficult to assess reading materials based on students' vocabulary size. Furthermore, it requires some technological skills to sort out all the words used in a text. In the present study, we attempted to create a way in which a high frequency word list, which is comprised of several sub-word lists with words of different characteristics, is used to profile specific texts written in English for use in language classes for bioscience

*生命科学部 EFL 研究室、**聖マリアンナ医科大学

majors.

As Laufer (1989) suggests, if an ESL (English as a Second Language) learner knows 95% of the words in a text, she can comprehend the text without much trouble. Obviously, this claim assumes that the learner has already acquired the basic grammar of English. However, at the same time, it is important to point out that without basic lexical knowledge most other language skills cannot be fully exploited during comprehension or production. Thus, the lexical knowledge of a learner may not be the most important aspect of language learning, but having sufficient vocabulary is crucial for post-beginning level learners. Based on Laufer's claim, it is hypothesized that the percentage of known vocabulary in a text indicates the difficulty of the text for the learner. If the text is entirely composed of known vocabulary, it becomes a simpler reading task for the learners, but if their lexical knowledge covers only 70 percent of the entire text, the readability may drop dramatically even with sophisticated top-down processing skills and background knowledge. The question is how accurately English instructors can predict the readability of texts for a specific group of learners.

Vocabulary profiling is one means of readability assessment, and at present there are several computer-based vocabulary profilers (e.g., Cobb 2007, Someya 2008) available for English educators. An adequate profiler can be a useful tool to facilitate the process of text selection, but available profilers are normally based on vocabulary extracted from a large general corpus, and therefore they are not sensitive enough to profile the use of specialized vocabulary in biosciences. Moreover, they do not take into consideration the vocabulary size of second language (L2) learners in a foreign language situation.

In order to fulfill the need for profiling authentic texts to be used in English programs for bioscience majors, it is necessary to create a good profiler. The profiler should be based on a word list that contains a sufficient proportion of technical and sub-technical scientific vocabulary, the basic vocabulary that most college students already know, and high-frequency general and academic vocabulary that are widely used in English. By using such a word list, it becomes possible to profile any type of text according to its easiness or difficulty and also its technicality. For this purpose, the word list should contain the following sub-lists:

1. Vocabulary that most college-level students already know
2. Widely used academic vocabulary
3. Widely used general vocabulary
4. Technical and sub-technical vocabulary used in Biosciences

The words in these sub-lists should be mutually exclusive. However, many words may belong to two or more sub-lists, so we have set a rule of priority, in which more basic vocabulary has higher priority, i.e., Basic -> Simple -> Moderate -> Academic -> General -> Life Science. The entire list is based on word families rather than lemmas. Lemmatization here means grouping lexical entries, inflected forms for verbs and plural forms for nouns, based on their dictionary forms rather than treating them as different words. A word family, on the other hand, contains not only the lemma list items but also derived words. Some word lists such as the Academic Word List (Coxhead 2000) and the General Service List (West 1953) categorize words by word families. Other word lists such as the JACET 8000 use lemmas. Because the aim of the present study is to find a way to profile texts for reading skill courses, we have decided to use word families. It is assumed that the meanings of derived lexical items can be inferred from the text if other forms are already known.

The present study consists of two phases: first we compiled a corpus of life science texts (hereafter referred to as the Life Science Corpus or the LSC) and a learner corpus (hereafter referred to as the LS Learner Corpus) in order to generate a high frequency word list for each corpus, and create a word list composed of several sub-lists. Secondly, we profiled various types of texts and small corpora using the word list (hereafter referred as *the LS Wordlist*[1]). By using *the LS Wordlist*, individual texts as well as corpora can be profiled using the same process. By profiling a corpus, we can identify some general tendencies of vocabulary use in a specific genre, and by profiling texts we can estimate the difficulty and technicality of each text. In this study we attempted to verify the usefulness of *the LS Wordlist*, which we have empirically generated, by profiling various types of texts and corpora.

Establishment of Corpora and High Frequency Word Lists

Two corpora, the LS Corpus and a learner corpus, and three general word lists, the Academic Word List (Coxhead 2000), the BNC Word List (Scott 2008), and the JACET 8000, were used to create the LS Wordlist. First, we compiled two corpora, the Life Science Corpus and the LS Learner Corpus, to be used for extracting two types of word lists. In order to compile the Life Science Corpus, we chose texts from 10 fields in the life sciences based on the curriculum offered at Tokyo University of Pharmacy and Life Sciences. For each field, we collected 50 texts of 2000 words (500 texts total) from textbooks, research articles, protocols and general science reading materials. From this one-million-word Life Science Corpus, we grouped words into word families and listed them in order of frequency. The LS Learner Corpus consists of 600 essays written by 400 undergraduate students at Tokyo University of Pharmacy and Life Sciences (200,000 words) during the first-year and second-year English language courses. We also grouped the words in this corpus into word families and listed them in the frequency order.

After creating two lists of words based on their frequency in two corpora, we generated a word list, *the LS Wordlist* (see *Table 1*), using the Academic Word List, the BNC Word List and the JACET 8000. *The LS Word* List is comprised of six sub-lists that reflect various types of vocabulary. The six sub-lists are *Basic, Simple, Moderate, Life Science, Academic* and *General* Wordlist. These are mutually exclusive.

The Basic Wordlist (1030 words) is based on the LS Learner Corpus. From this word list, we selected words that appeared more than 50 times in this corpus. We have further divided *the Basic Wordlist* into five levels based on frequency: the Basic-100 (1-100), Basic-200 (101-200), Basic-300 (201-300), Basic-500 (301-500), and Basic-1030 (501 to 1030). The LS Learner Corpus is a corpus of learners' production, and some words that the students know but do not use in their writing may not have been included. The literature (Webb 2008) reports that there is a gap between learners' productive and receptive vocabulary

[1] In order to distinguish the word list we created from other word lists, we use the term wordlist and italicize it to refer to our own word lists in this paper.

sizes. In order to fill this gap, we created additional lists of the students' known vocabulary. They are *the Simple Wordlist* (947 words) and *the Moderate Wordlist* (1092 words). We gave 10 students a word recognition test based on the JACET 8000 word list, and then selected the words that the majority of the students knew (*the Simple Wordlist*) and that half of them knew (*the Moderate Wordlist*). *The Life Science Wordlist* (1178 words) was generated from the LSC. From the one-million-word LSC, we selected frequently used words that covered roughly 95% of the entire LSC, and then all the words that belong to *the Basic, Simple* and *Moderate* lists were removed.

In addition to the above sub-lists, we made *the General Wordlist* (80 words) and *the Academic Wordlist* (246 words) based on the BNC Word List (Scott 2008) and the Academic Word List (Coxhead 2000). These two were generated in order to include essential general and academic lexical items that do not appear in other sub-lists.

The LS Wordlist (4573 words)						
Sub-lists	*Basic* (5 levels)	*Simple*	*Moderate*	*Life Science*	*Academic*	*General*
# of words	1030	947	1092	1178	246	80
Sources	Learner Corpus	WR Test	WR Test	LS Corpus	BNC WL Academic WL	BNC WL Academic WL

Table 1. The Components of the LS Wordlist

Profiling

Using *the LS Wordlist*, we profiled two types of materials, corpora and individual texts. We first profiled corpora and then profiled individual texts from various sources. Three commercial software products, *WordSmith Tools v.4, Microsoft Excel 2004 for Macintosh and FileMaker Pro v.7* were used. The following is the basic procedure:

1. Make a word-family based word list of the corpus (or text) to be profiled, using *Wordsmith Tools v.4.*
2. Match the word families with the ones on *the LS Wordlist* using *FileMaker Pro v.7.*
3. Group the word families based on the sub-lists.
4. Add up the frequency counts of word families in each sub-list.

The output can be interpreted based on the coverage of each sub-list. If a text contains more vocabulary that belongs to the *Basic* and *Simple* sub-lists, then the text is supposedly easier than a text that contains fewer words on those sub-lists. If the text contains more words that belong to the *Life Science* sub-list, then the text is more technical / scientific.

By profiling corpora, the characteristics of vocabulary use in a particular genre become evident. In the present study, we profiled three different types of corpora: newspaper English, scientific research articles and textbooks. We compiled a corpus of scientific newspaper articles, a corpus of scientific research articles and a corpus of scientific textbook passages for this purpose. The newspaper article corpus (860,000 words) is a collection of 1168 scientific articles from various newspapers including *International Herald Tribune, Washington Post, USA Today*, and *The Japan Times* and *The Daily Yomiuri* from January to August in

2006. For the research article corpus (400,000 words), we collected 89 scientific reading materials used in freshman seminars at Tokyo University of Pharmacy and Life Sciences. We also examined four textbooks used in English classes at St. Marianna University School of Medicine and generated a medical corpus. The textbooks are about anatomy and physiology of human body including basic medicine as well as basic clinical medicine. All the textbooks target students of universities and medical schools in the United States.

Fig. 1 shows the profiles of the three corpora. The newspaper article corpus contained more basic words than the other two corpora, which suggests the relative ease of the texts in newspaper articles in general. Because this is a corpus of newspaper science articles, a significant portion of technical vocabulary categorized as LS vocabulary was present. The research article corpus contained a larger portion of the LS and Basic-1030 vocabularies. The biggest difference between these two corpora was the proportion of LS vocabulary and less frequent basic words. The coverage of *the LS Wordlist* as a whole was only 70% or so in the medical corpus, and there was a large proportion from the category "other." Although a number of different words such as proper nouns, acronyms, and infrequent vocabulary may be included in this category, it was assumed that a large part of it is specialized vocabulary used in medicine.

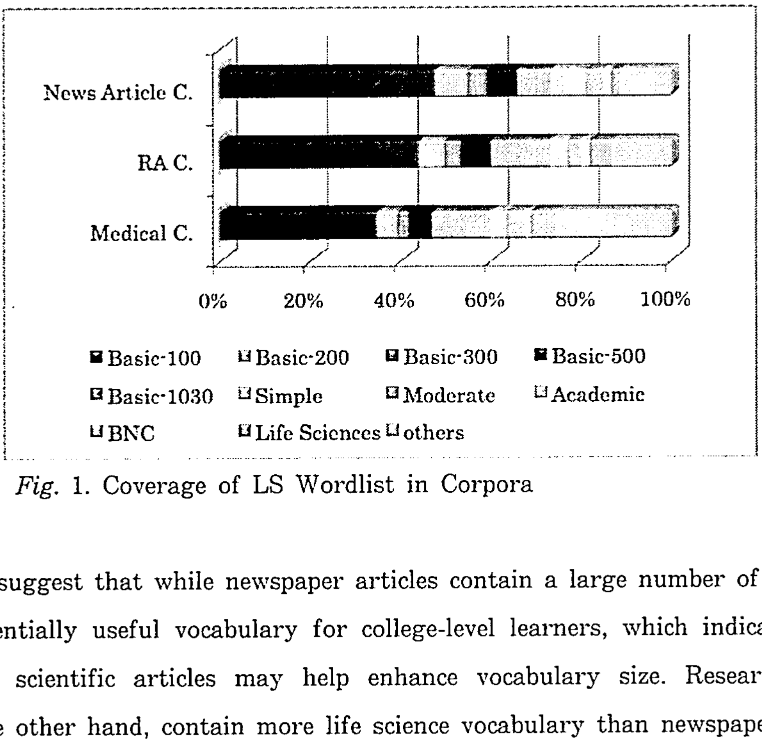

Fig. 1. Coverage of LS Wordlist in Corpora

This results suggest that while newspaper articles contain a large number of known vocabulary, they also contain potentially useful vocabulary for college-level learners, which indicates that for vocabulary building reading scientific articles may help enhance vocabulary size. Research articles and medical textbooks, on the other hand, contain more life science vocabulary than newspaper articles. This suggests that reading research articles may help learners acquire more specialized vocabulary, but the fact that the medical corpus contains a large number of words in the "other" category means that textbooks written for medical students in English speaking countries may generally be too difficult for students in an EFL (English as a Foreign Language) program to read. In order to fully comprehend those texts, a wider range of vocabulary is needed. Unlike using newspaper articles, in order to use those medical and scientific texts in an EFL classes, it is necessary to profile individual texts and carefully single out appropriate ones.

To investigate whether *the LS Wordlist* is a good profiler for analyzing individual texts, we profiled several scientific article; excerpts from an English for academic purposes textbook; excerpts from

Harrison's Internal Medicine, which is one of the required books for medical school students; a TV drama script dealing with a medical situation; a juvenile novel, *Anne of Avonlea*; and Darwin's *On the Origins of Species*. Scientific articles were taken from articles targeted at general readers as well as researchers such as *Scientific American*.

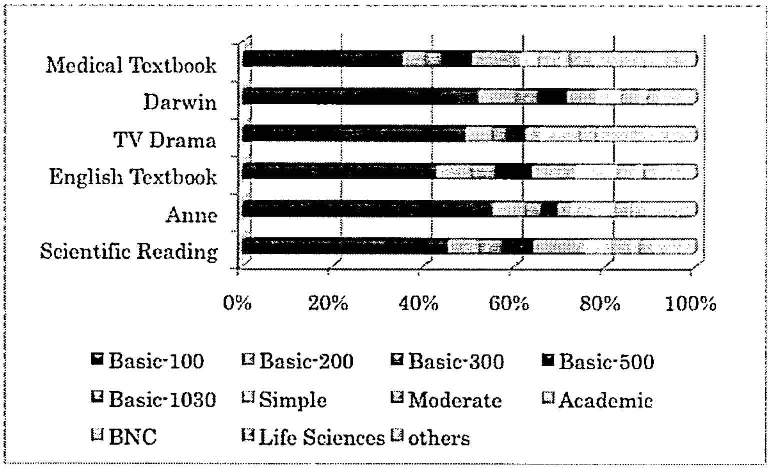

Fig. 2. Coverage of the LS Wordlist in text samples

Fig. 2 shows the results for the individual texts. Over 50% of the vocabulary *Anne of Avonlea* contains is comprised of the 100 easiest word families that every college student in Japan should know. Darwin's On the Origins of Species also shows a similar pattern, but it contains a portion of life science vocabulary. The TV drama script is the only spoken language data in this sample, and it shows a different pattern compared to other texts. There is a small portion of both academic and life science vocabulary in this drama script. This suggests *the LS Wordlist* we have created cannot fully analyze spoken language data, and this is also true for the medical textbook sample. Both texts contain a large portion of "other" vocabulary. In these texts such as TV drama scripts, there is a large proportion of proper names. Medical texts may also contain a large number of names such as the names of chemicals, enzymes, diseases, etc. Proper names can be treated as known vocabulary in many cases, but since *the LS Wordlist* does not contain a large list of proper names, this contributed to a higher number of words categorized as "other," making it difficult to determine the difficulty and technicality of these texts at this stage. The vocabulary in the "other" category may include biomedical terminology, such as names of chemicals or disorders, and anatomical as well physiological terms. It is necessary to analyze those vocabulary items to find out whether it is necessary to make an additional sub-list for medical English.

It seems *the LS Wordlist* can be used to profile texts written for the general public with considerable sensitivity to both general and technical vocabulary. However, in order to profile texts written for professionals in the bioscience field, a larger list with more technical vocabulary is necessary. For the use in undergraduate English language courses in Japan, *the LS Wordlist* can be a useful tool for profiling.

Limitations and Suggestions for Future Studies

A word list with different types of words may be a good profiling tool in some ways, especially for choosing suitable teaching materials. Since EFL students in a specific academic field are/will be exposed to vocabulary in their particular field more extensively, profiling based on vocabulary types provides richer information for teachers as well as for learners. Profiling texts using *the LS Wordlist* is potentially very useful, but the present list cannot deal with specialized medical vocabulary and proper nouns, both technical and non-technical. Also *the LS Wordlist's* consistency has to be reviewed again. In order to do this, a larger corpus may be necessary to generate more reliable word lists for profiling purposes. In addition, verification of basic word lists would need to be done since individual students may have different ranges of vocabulary. Further refinement is definitely needed before we can use this profiler for practical applications. But, even with *the LS Wordlist* alone, we can observe the characteristics of small corpora and individual texts.

References

Cobb, T. (2007). *The Compleat Lexical Tutor.* v. 6.2. Retrieved July 3, 2008, from http://www.lextutor.ca/

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly*, 34(2): 213-238.

Coxhead, A. and Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew (Ed.) *Research perspectives on English for Academic Purposes.* (pp. 252-267). Cambridge: Cambridge University Press.

Darwin, C. (1872). *On the origin of species.* 6th edition. Retrieved from http://www.gutenberg.org/etext/2009.

Dudley-Evans, T. and St. John, M. (1998). *Developments in English for specific purposes.* Cambridge: Cambridge University Press.

Frase, S. (2002). A statistical analysis of the vocabulary of medical research articles (3): Technical and subtechnical vocabulary. *Integrated Studies in Nursing Science.* 4(2): 27-45.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (Eds.), *Special language: From humans thinking to thinking machines.* (pp. 316-323). Clevedon: Multilingual Matters.

Montgomery, L. M.(1909). *Anne of Avonlea.* Retrieved from http://www.gutenberg.org/files/47/47-h/47-h.htm.

Nation, I.S.P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Scott, M. (2008). The BNC wordlist. Retrieved October 10, 2007, from http://www.lexically.net/wordsmith/ index.html.

Someya, Y. (2008). *Word Level Checker.* Retrieved September 22, 2007, from http://www.cl.aoyama. ac.jp/english/newSite/wlc/index_J.html.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(1): 79-95.

West, M. (1953). *A general service list of English words.* London: Longman, Green & Co.