Corpus Research and Its Application in ESP Programs
Lexical Profiling of Reading Materials Using Frequency Vocabulary Lists

Akiko Hagiwara*, Mao Naito**
*School of Life Sciences, Tokyo University of Pharmacy and Life Sciences,
1432-1Horinouchi, Hachioji-shi, Tokyo, Japan
hagiwara@ls.toyaku.ac.jp
**English Program, St. Marianna University School of Medicine,
2-16-1 Sugao, Miyamae-ku, Kawasaki-shi, Kanagawa, Japan.
m2naito@marianna-u.ac.jp

The present study explores the use of technology to profile authentic reading materials in the biomedical field, which can be used in both traditional classrooms and computer-assisted learning environment. In order to accommodate the needs of students majoring in biomedical sciences, which require lexical knowledge specific to the field, using content-based reading materials not only enhances students’ vocabulary size but also helps them acquire state-of-the-art information in a fast moving field. Profiling reading texts using a vocabulary list containing frequently used vocabulary in biomedical literature facilitates the process of text selection in ESP programs. Based on several corpora a word list (LS Wordlist) comprised of several sub-lists reflecting various types of vocabulary was created. Three small corpora and nine texts were profiled based on this word list. The output clearly indicates the characteristics of texts, showing how easy or technical the text is for the learners.

1. 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 style (Dudley-Evans and St. John, 1998) and makes use of specific vocabulary and widely used sub-technical vocabulary (Fraser 2001). However, controlling the vocabulary in text selections is a challenging task, because reading materials in ESP programs need to contain a sufficient number of specialized terms including both technical and sub-technical vocabulary, but they should also be readable by the learners.

As Laufer (1989) suggests, if a reader knows 95% of the words in a text, she can comprehend it without much trouble. Based on this claim it is hypothesized that the percentage of known vocabulary in a text can predict the difficulty of it for the learner.

Vocabulary profilers can be a useful tool to facilitate the process of text selection, but available profilers (e.g., Cobb 2007, Someya 2008) that are based on vocabulary used in general texts do not suffice for the needs of ESP in the biomedical field. Moreover, they do not take into consideration the vocabulary size of second language (L2) learners in a foreign language situation.

The present study attempted to profile texts and small corpora using a word list comprised of six sub-lists. By using this word list, individual texts as well as corpora can be profiled in the same way. We can obtain some general tendency of vocabulary use in a specific genre by profiling a corpus, and we can estimate the difficulty and technicality of a single text by profiling a text. In this study we attempted to test the usefulness of the word list we have empirically generated by profiling various types of texts and corpora.

2. Methods
First, we generated a word list, LS Wordlist, comprised of six sub-lists that reflect various types of vocabulary. Since this study aims at selecting reading materials, we have decided to categorize words into word families rather than lemmas. 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 a learner corpus consisting of 600 essays written by 400 undergraduate students (200,000 words). We made a frequency word list and selected words that appeared more than 50 times in this corpus. In order to supplement the data, we created additional lists of 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 JACET 8000 word list. We selected the words that the majority of the students knew (the Simple Wordlist) and that half of them knew (the Moderate Wordlist). We have further divided the Basic Wordlist into five levels based on frequency. Those are Basic-100 (1-100), Basic-200 (101-200), Basic-300 (201-300), Basic 500 (301-500), Basic 1030 (501 to 1030).

The Life Science Wordlist (1178 words) is generated from the Life Science Corpus (LSC). In order to create LSC, we chose texts from 10 fields in the life sciences based on the curriculum offered at the Tokyo University of Pharmacy and Life Sciences. For each field, we collected 50 texts of 2000 words (total 500 texts) from textbooks, research articles, protocols and general science reading materials. From this one-million-word LSC, we grouped words into word families, and we selected frequently used words that covered 95% of the entire LSC.

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

By the use of three commercial software products, WordSmith Tools v.4, Microsoft Excel 2004 for Macintosh and FileMaker Pro v.7, we profiled three small corpora and six individual texts based on the LS Wordlist.

The Newspaper article corpus (860,000 words) is a collection of 1168 scientific articles from various newspapers between January and 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 textbooks used in English classes at St. Marianna University School of Medicine and generated a medical corpus. These textbooks are all targeted at the students of universities and medical schools in the United States.

As for the texts, we profiled scientific readings targeted at general readers as well as researchers, an English textbook for academic purposes, and a medical text for internal medicine. In addition, Anne of Avonlea and the Origin of Species were also examined.

3. Results and Discussion

Regarding corpora, as Fig. 1 shows, while 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, there were some technical words categorized as LS vocabulary.

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 of the category “Others.” Although there are number of different items such as proper nouns, acronyms, and infrequent vocabulary included in this category, it was assumed that a large part of it is specialized vocabulary used in medicine.

Figure 2 shows the results for the individual texts. Anne of Avonlea contains 50% of vocabulary comprised of the 100 easiest word families that every college student in Japan should know. Darwin’s 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 which is also true for the medical textbook sample. Both texts contain a large portion of “Others” vocabulary. Unless we further analyze the data in the “Others” category, it is difficult to determine the difficulty and technicality of these texts.

4. Limitations and Suggestions for Future Studies

Profiling texts using the LS Wordlist is potentially very useful, but the present LS Wordlist cannot deal with specialized medical vocabulary and proper nouns, both technical and non-technical. Further refinement is definitely needed before we can implement this profiler into a practical application, but even with this LS Wordlist, we could visually see the characteristics of small corpora and individual texts.

References