Automatic Generation of High-Frequency Vocabulary Lists and Level-Adjusted Exemplar Sentences for Non-Native Speakers of English

アントニ　ローレンス

早稲田大学理工学術院英語教育センター　〒169-8555 東京都新宿区大久保 3-4-1
E-mail:　anthony@waseda.jp

概要

過去研究によると第 2 言語学習者にとって高頻度の単語の Direct Learning が効果的である。また、高頻度の単語に関して語彙知識の深さ（Depth）も重要である。この結果から高頻度の単語の学習に当たって、単純な L2-L1 語彙リストは明らかに不十分な学習道具である。高頻度の単語を学習する際、語彙の語源、派生、活用などに直面すべきであり、意味を持つ典型文も目にする必要がある。また、その典型文に含まれている語彙が適切な語彙レベルであることも 1 つの条件である。この論文では新しい語彙リストと典型文作成ツールを提案する。このツールにより、語彙アイテム、発音ガイド、適切な語彙レベルでの典型文などを自動的に作成し、すぐに印刷し、学生に配布できる形にアウトプットする。ツールの機能と設定が柔軟であり、生成されたリストも既に日本における大型の理工系英語教育プログラムで使用され、効果的な語彙学習教材であることが証明されている。

Automatic Generation of High-Frequency Vocabulary Lists and Level-Adjusted Exemplar Sentences for Non-Native Speakers of English

Laurence ANTHONY

Waseda University, Faculty of Science and Engineering, Center for English Language Education
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
E-mail: anthony@waseda.jp

Abstract

Previous research on vocabulary learning has shown that non-native speakers of English benefit greatly from the direct learning of high-frequency vocabulary items. It is also known that increasing the depth of the learners' knowledge of high-frequency items is important. These findings suggest that simple L2-L1 vocabulary lists are insufficient. Instead, learners should be exposed to high-frequency items in a variety of forms and contexts, together with meaningful exemplar sentences that are also at an appropriate vocabulary level. In this paper, I present a novel vocabulary list and exemplar sentence generator that automatically creates an attractively formatted table of vocabulary learning items, pronunciation guides, and exemplar sentences at an appropriate learner level. The generator is highly customizable and the lists produced by the tool have already proved successful in a very large English program for students of science and engineering in Japan.

1. Introduction

Vocabulary instruction has traditionally been a central component of English language courses. In recent years, with new tests of vocabulary knowledge and the creation of large corpora of general and specialized English, our understanding of vocabulary and its importance in English as a Foreign Language (EFL) instruction has increased dramatically. Today, applied linguists know much more about what vocabulary items are important to learn and how vocabulary items at different frequency levels should be taught. On the other hand, implementing these ideas in a real-world EFL program has proven to be more problematic. For example, although it is now clear that students can be successfully taught high-frequency vocabulary items using a direct approach [1], providing a contextual link between such lists of high-frequency words and other learner materials

アントニ　ローレンス, "Automatic Generation of High-Frequency Vocabulary Lists and Level-Adjusted Exemplar Sentences for Non-Native Speakers of English"
日本英語教育学会第 41 回年次研究集会論文集, pp. 7-14, 日本英語教育学会編集委員会編集, 早稲田大学情報教育研究所発行, 2012 年 3 月 31 日.
This proceedings compilation published by the Institute for Digital Enhancement of Cognitive Development, Waseda University.

can be challenging. Also, it can be assumed that learners require a greater depth of knowledge of high-frequency words than they do for low frequency words. This implies that they are exposed to high-frequency words in varying forms (i.e., inflections and derivations) and contexts. Again, however, providing suitable materials that develop vocabulary depth can be difficult.

In this paper, I will first briefly review the current theories on vocabulary learning and discuss the importance of high-frequency vocabulary items. Next, I will present a novel tool for creating vocabulary materials that addresses the current challenges facing EFL materials developers. The system presented here takes as input a pre-defined list of vocabulary items, an upper bound of the user's general vocabulary level, and a set of electronically stored course texts that serve as a source of exemplar sentences. Usually, the pre-defined vocabulary lists will be composed of high frequency general words that are selected from traditional vocabulary resources or generated dynamically using corpus tools. However, the lists might also be more specialized, such as lists of technical vocabulary used in specialist disciplines. The upper of bound of the user’s general vocabulary, on the other hand, will be defined based on known research estimates or measured using a standard vocabulary level test.

Using this input, the tool then automatically identifies which target items appear in the course texts and presents these in an attractively formatted table together with frequency-level information, pronunciation guides, and carefully selected exemplar sentences from the course materials that are matched to the user's general vocabulary level.

The tool has already been successfully applied in the creation of vocabulary lists for a very large English program for students of science and engineering in Japan. A brief overview of the program will be given and the vocabulary lists created by the tool will be discussed. Finally, possible improvements to the tool will be suggested.

2. Review of Literature on Vocabulary Learning

Early studies on the vocabulary size of native speakers suggested that adult native speakers have at their disposal a huge amount of vocabulary that would be almost impossible for foreign language learners to acquire (for a review of this research, see [2]). However, when researchers began investigating the size of native speakers' vocabulary knowledge more exactly, a surprising result was found. Nation [1], for example, showed that native speakers typically learn only 1000 word families each year of their life. In this work, Nation defined a word family as a base word (e.g. "organize") together with its inflections and close derivations (e.g., "organizing," "organized," "organization"). Nation's estimate was also supported by Beck and McKeown [3], who calculated that children aged five to six know only around 2500 to 5000 words. From these results, Waring and Nation [4] estimated that a typical university graduate would know around 20,000 word families. This is, of course, much smaller than the number of words in a typical dictionary. For example, Webster’s 3rd International Dictionary contains around 54,000 word families [4].

Similar research has also been carried out to determine the typical vocabulary size of non-native speakers of English. In the same paper as above, Nation & Waring [4] reported that many adult learners of English as a foreign language know less than 5000 word families. A more recent study by Chujo [5] also suggests a similar figure, with Japanese university students demonstrating a knowledge of only around 3000 to 5000 words. In addition, researchers have also identified a problem in the complexity level of vocabulary items that EFL learners have typically acquired. Rather than acquiring the most frequent words used by native speakers, many non-native learners of English have been shown to have acquired some very low-frequency words that they are unlikely to encounter in real-life situations [6]. One reason for this is perhaps the wash-back effect on high school education programs of university entrance exams that tend to include very difficult English reading passages [6].

The importance of high-frequency words over low frequency words is clearly revealed when the relationship between frequency and coverage is plotted. Table 1 and Figure 1 show the coverage of the most frequent words in the Brown Corpus plotted for the top five 1000-word levels [7]. The results show that knowledge of the first 1000 words will give learners a coverage of 72% of all the words in the corpus. Knowing an additional 1000 words, on the other hand, will only improve that coverage by 7.7%, and a further 1000 words, will only improve on that by 4.3%. Clearly, if a non-native speaker hopes to understand English, knowledge of the most frequent words is essential.

Table 1. Relationship between Vocabulary Size and Coverage in the Brown Corpus

Brown Corpus (100 million words)		
Vocabulary Level	Coverage (%)	Cumulative Coverage (%)
1000	72	72
2000	7.7	80
3000	4.3	84
4000	2.8	87
5000	1.9	89

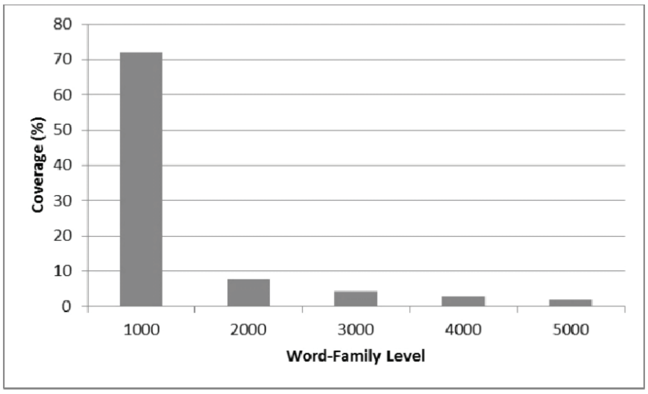


Figure 1. Relationship between Vocabulary Size and Coverage in the Brown Corpus

What is not clear from Table 1 or Figure 1 is how many frequent words a learner would need to comprehend a text without support materials, such as dictionaries, glossaries, or native-speaker informants, assuming that some words could be understood through context. However, this question was investigated by Laufer [8], who found that a coverage of 95% of the running words was needed. Later, research by Nation [1], suggested that this value may even be as high as 98%. Of course, not all the words resulting in a 95% or 98% coverage will have a high frequency; highly specialized texts in science and engineering, for example, are likely to include many words that have a low frequency in general English. Nevertheless, the importance of exposing learners to high-frequency words is clearly a beneficial first step to achieving this 95% or 98% coverage level.

A related issue is the kind of information learners should be expected to know about vocabulary at different frequency levels. Nation [1:27] lists multiple facets of vocabulary knowledge that can be important. These are the sound of the word, its written form, its constitute word parts, its core meaning, the concepts it is associated with, other words associated with it, the grammar patterns that it is likely to appear in, the words it is likely to co-occur with, and the registers and genres that the word is likely to be used in. When combined, these facets serve as a measure of vocabulary *depth* (in contrast to the sheer number of words known, which is usually referred to as *breadth*). Other researchers have investigated the relationship between depth and breadth and there is evidence to suggest that the two are highly correlated (e.g., [9], [10]).

Although several researchers (e.g., [11-13]) have argued that English has a monosemic bias, i.e., words have a single, fundamental meaning, it is clear that the depth of knowledge needed to fully 'know' a word will still be related to the word's frequency level. High-frequency words, for example, almost by definition, will occur in more varied contexts, have more concepts associated with them, and be used in a greater number of registers and genres than low frequency words. There is also a strong argument that high-frequency words will exhibit multiple core meanings. A simple example is the very high-frequency word "bank," which can refer to an establishment that stores peoples savings or the side of a river. It can appear as a noun or a verb, and will be a frequent term in both academic and non-academic texts. In contrast, the very low frequency word "supermacroporous" is likely to only occur as an adjective and appear only in the narrow genre of academic research papers in the field of materials science.

The examples of "bank" and "supermacroporous" strongly demonstrate that learners will generally require a greater depth of knowledge to master high-frequency words and a greater breadth of knowledge if they hope to master low frequency words of many varying disciplines, genres, and registers.

A separate strand of research has focused on the best way for learners to acquire both high- and low-frequency vocabulary items. Contrary to beliefs among some teachers, research has shown that a direct approach can be successful, especially when dealing with high-frequency words (e.g., [1], [16]). However, this does not imply that direct approaches are the only approach that should be used. On the contrary, Nation [1] clearly states that both direct and indirect approaches should be seen as complimentary. Also, he argues that direct approaches should only account for a small percentage (25%) of the time allocated to the vocabulary component of a language program. Also, as the need to learn increasingly less frequent words, such as technical vocabulary in science and engineering, arrives, the use of incidental learning approaches becomes preferred.

The above review of literature suggests several key features of a successful vocabulary program that are relevant to the current study: 1) it should provide students with materials to help them master the most frequent words of English through both direct and incidental learning methods, 2) it should provide learners with information to develop their depth of knowledge of high-frequency vocabulary items, and 3) the materials should facilitate both direct and incidental learning approaches. One final point is that vocabulary learning is only one component of an effective language program. Therefore, the time necessary to generate vocabulary lists should be minimized to allow teachers to focus on other aspects of teaching preparation.

This leads to the following question: Is it possible to automate the creation of effective lists for the learning of high-frequency vocabulary items that lend themselves to direct teaching approaches whilst also providing suitable information to improve vocabulary depth knowledge acquired through both direct and indirect learning approaches. In the following section, I introduce *AntLister*, a novel, automated, vocabulary list and exemplar sentence generator that demonstrates that such a vocabulary list creation tool can be developed with today's computer hardware and software.

3. Design and Development of *AntLister*

AntLister is a software tool designed to automatically create vocabulary lists and exemplar sentences. The software is built using the Perl 5.14 programming language and runs on the command line of a Windows operating system. For a specified target word, the tool is programmed to generate the following information: a) a numbered index used for in-class instruction, and review purposes, b) the target word itself, c) an IPA pronunciation guide, d) an L1 translation, e) the position of the first occurrence of the word in a set of EFL course materials, f) the frequency level of the word, g) one or more exemplar sentences from the source materials that demonstrate varying uses of the word, and g) one or more exemplar sentences from an alternative reference source.

To automatically generate an appropriate IPA pronunciation guide, the Carnegie Mellon University Pronouncing Dictionary is utilized [17]. This is an electronic dictionary of 125,000 words, each with an Arpabet phonetic transcription code [18]. Arpabet is a phonetic transcription code developed by the Advanced Research Projects Agency (ARPA) in the 1970s for use in speech recognition systems, where each phoneme is represented by a distinct sequence of ASCII characters, e.g., green (G R IY N), she (SH IY), and yield (Y IY L D). By programming a simple mapping from the Arpabet phonetic transcription code to IPA symbols, an automated IPA pronunciation guide could be created.

To provide an automated L1 (Japanese) translation of the target word, data from the EDICT (Japanese/English Dictionary Project) is utilized [19]. In cases where multiple L1 translations are possible, the first word in the list dictionary entry is provided. However, in some cases no direct L1 translation is available and in these cases the entry is left blank.

To identify the first occurrence of the target word in the EFL course materials, an electronic version of the materials is stored on disk with each unit of the materials saved as a separate file. Then, a simple software scan reveals in which file, and hence in which unit, the target word appears.

The creation of a set of exemplar sentences highlighting varying inflections and near-derivations of the target word is achieved in a pipe-line process. First, 16 baseword lists created by Paul Nation [20] are inputted into the system. These consist of the most frequent 16,000 word families in the British National Corpus grouped under head words and composed of a set of 'family members,' which correspond to inflections and near-derivations of the head word. For each target word, all baseword lists are searched and the associated family members of the target word extracted. Next, the system searches in the supplied course materials to find the first occurrence of each family member together with its surrounding context (i.e., the sentence in which it occurred). If no occurrence is found, the family member is deleted. Similarly, the system also searches for ‘hit’ sentences in a reference corpus from the specialist subject of the learners. This results in the final exemplar sentences containing useful specialist-subject words that can be learned incidentally while learners examine the usage of the target words.

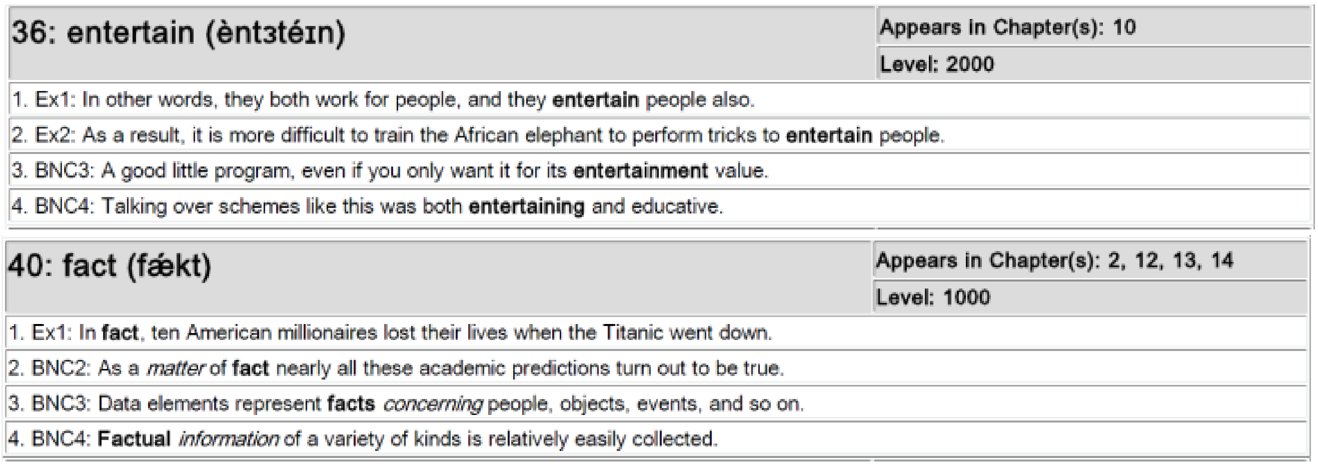
Next, all 'hit' sentences are subject to a high-pass filter that evaluates their vocabulary difficulty level by comparison with a pre-defined user vocabulary level. If the sentences contain less than a pre-defined coverage level of words assumed to be known by the user (e.g., 95%), they are removed. Finally, the remaining exemplar sentences are randomized and a predefined number of them added to the vocabulary list. If the remaining sentences contain words that highly collocate with the target word, these collocating words are also marked.

To present the results to the user in an easy-to-understand format, the resulting information is combined into a dynamically created HTML formatted documented (and associated CSS style sheet) and output to a Web browser display for viewing. For distribution to learners, the HTML document can then be saved as a PDF file.

4. Results and Discussion

An example output from *AntLister* is shown in Figure 2. Note that in this case, no L1 translation has been included. Also, note that two sources of exemplar sentences have been incorporated; Ex. 1 and Ex. 2 refer to examples from course materials, whereas BNC3 and BNC4 refer to examples taken from a secondary source, i.e., the science section of the British National Corpus.

The table of vocabulary and exemplar sentences created by *AntLister* is created in a completely automated fashion. Not surprisingly, therefore, some of the information in the table can be



36: entertain (èntɜtéɪn)

Appears in Chapter(s): 10

Level: 2000

1. Ex1: In other words, they both work for people, and they **entertain** people also.
2. Ex2: As a result, it is more difficult to train the African elephant to perform tricks to **entertain** people.
3. BNC3: A good little program, even if you only want it for its **entertainment** value.
4. BNC4: Talking over schemes like this was both **entertaining** and educative.

40: fact (fǽkt)

Appears in Chapter(s): 2, 12, 13, 14

Level: 1000

1. Ex1: In **fact**, ten American millionaires lost their lives when the Titanic went down.
2. BNC2: As a *matter* of **fact** nearly all these academic predictions turn out to be true.
3. BNC3: Data elements represent **facts** *concerning* people, objects, events, and so on.
4. BNC4: **Factual** *information* of a variety of kinds is relatively easily collected.

Figure 2. Output from *AntLister* for the words "entertain" and "fact"

problematic. One of the early problems was creating an acceptable IPA conversion. A one-to-one conversation of Arpabet phonetic transcription codes to IPA symbols was unacceptable. Following the advice of an experienced phonetician, the solution was to map codes and in some cases code-sequences directly to IPA symbols.

Another problem was the selection of suitably meaningful exemplar sentences. If the required coverage percentage was set at 98% or above, following the research findings of Nation (2001), the high-pass filter tended to remove sentences that provided contextual information and details. The result was sentences that were easy to read but rather abstract. On the other hand, if the coverage percentage was set lower, e.g., below 95%, the remaining sentences would become meaningful but overly complex. The impact of the coverage parameter is shown in Table 2. Based on preliminary experiments, a filter setting of 95% was set as the default.

Table 2. Impact of coverage percentage setting on the *AntLister* high-pass filter sentence selector

Coverage setting	Resulting sentences from *AntLister*
90%	Additionally, the offline manager must respond to any mail sent regarding failures or unavailability of media items or units.
95%	If the information is preserved, it will be in an effort to guarantee its availability in case of legal dispute.
98%	Attendance at lectures does not appear to be affected by the availability of outline notes.
99%	Firstly, it is not clear that data are unavailable for some of the countries.

To test *AntLister* in a real-world environment, the tool was used to create vocabulary lists for two required first-year courses in the Waseda University CELESE English program for scientists and engineers. Details of the CELESE English program can be found elsewhere [21], but it is important to note that the program is a very large-scale English program serving 10,000 undergraduate and graduate students in Japan. All first-year students (approx. 2000 in total) are required to study courses in *Communication Strategies* and *Academic Listening Comprehension*. Both courses extend across two semesters and have an important vocabulary component. Generally, both courses require that students develop their vocabulary knowledge outside of the classroom as part of a self-study component. On the other hand, some teachers of the courses include explicit vocabulary instruction in class. Therefore, the vocabulary materials needed to be flexible enough to facilitate both teacher-led and self-study learning approaches. Due to the extensive time it would take to develop vocabulary lists for these two courses, the CELESE program designers felt this would be ideal testing ground for *AntLister*. The full set of materials created for the program can be downloaded at the following website: http://www.celese.sci.waseda.ac.jp.

To date there has been no controlled study to measure the effectiveness of the automatically created vocabulary lists over traditional textbook or handmade lists. However, several indirect indicators of its success in the CELESE program are available. First, since introducing vocabulary through these automated lists, students in both courses have shown increasing gains on standardized tests, including the TOEIC, as shown in Figure. 3. These results suggest not only that students are using the lists successfully, but also that they are incidentally learning other important but lower-frequency specialist-subject words that are contained in the exemplar sentences.

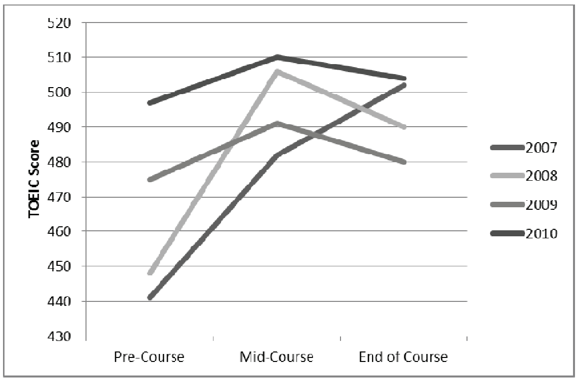


Figure 3. Performance Gains on TOEIC for First-Year Students at CELESE (2007 to 2011).

Second, while teachers tend to dismiss automated approaches to materials development, no complaints have been voiced by either the full-time or part-time faculty at CELESE regarding the vocabulary lists automatically created by *AntLister*. On the contrary, the success of the lists created for use in first-year courses has led to the creation of similar lists for 3rd- and 4th-year elective courses in technical writing and presentation.

Despite the positive results of the *AntLister* vocabulary lists, there are many areas for improvement. First, the lists do not always include the most meaningful examples. Clearly, there is a strong relationship between lexical richness and sentence detail, as shown by the sentences in Table 2. Therefore, the challenge is to identify sentences with a low level of vocabulary complexity but a high level of concreteness. Another problem is that each auto-generated vocabulary list entry can take several minutes to generate depending on the number of source materials that need to be searched. In the case of the CELESE materials, a complete list required several hours of processing time. Although this is still many times faster than an equivalent human process, a faster algorithm is certainly desirable. Finally, the current system runs on the command line of a Windows operating system. For wider applicability, the software needs to be compiled to run as a multiplatform, single-file executable with an intuitively-designed user interface.

5. Conclusion

This paper has presented a novel approach to the automatic creation of vocabulary lists for use in English as a Foreign Language (EFL) classroom contexts. The approach used here relies on a combination of online sources, including an English pronunciation dictionary, an English-Japanese translation dictionary, and a source of sentences that demonstrate target word usage in a variety of contexts. It also relies on a programming script that implements a set of filters to carefully select exemplar sentences that fulfill a set of 'good exemplar' criteria suggested by research. The approach has been applied in the creation of vocabulary lists for required and elective courses in the Waseda University CELESE English program for scientists and engineers, and teacher comments and student performance gains suggest that the approach is effective. However, there are still many improvements that can be made to the software, and a controlled study to measure the tool's effectiveness over traditional hand-created lists is needed. This is the aim of future work.

References

[1] Nation, I.S.P., Learning vocabulary in another language, Cambridge: Cambridge University Press, 2001.

[2] Nation, I.S.P., Vocabulary size, growth and use, in R. Schreuder and B. Weltens (Eds.), The Bilingual Lexicon. Amsterdam/Philadelphia: John Benjamins, 115-134, 1993.

[3] Beck, I.L. and McKeown, M.G., Social studies texts are hard to understand: Mediating some of the difficulties. Language Arts, 68, 482-490, 1991.

[4] Waring, R. and Nation, I.S.P., Vocabulary size, text coverage, and word lists, in N. Schmitt and M. McCarthy (Eds.), Vocabulary: Description, Acquisition and Pedagogy, Cambridge: Cambridge University Press, pp. 6-19, 1997.

[5] Chujo, K., Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list, in J. Nakamura, N. Inoue, T. Tabata (Eds), English Corpora Under Japanese Eyes, Language and Computers. Amsterdam: Rodopi, pp. 231-249, 2004.

[6] Browne, C. and Culligan, C., Combining technology and IRT testing to build student knowledge of high frequency vocabulary, The JALT CALL Journal, 4 (2), pp. 3–16, 2008.

[7] Francis, W.N. & H. Kucera, Frequency Analysis of English Usage: Lexicon and Grammar. Boston: Houghton Mifflin. 1982.

[8] Laufer, B., What percentage of lexis is essential for comprehension, in C. Lauren & M. Nordman (Eds.), From humans thinking to thinking machines, Clevedon, UK: Multilingual Matters, pp. 316-323, 1989.

[9] Qian, D.D., Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. Language Learning, 52, 513-536, 2002.

[10] Laufer, B. and Nation, I.S.P., Vocabulary size and use: lexical richness in L2 written production. Applied Linguistics, 16, 307-322.productive ability. Language Testing, 16, 33-51, 1995.

[11] Ruhl, C., On monosemy: A study in linguistic semantics. Albany: State University of New York Press, 1989.

[12] Parent, K., Polysemy A second language pedagogical concern, Unpublished doctoral dissertation, Victoria University of Wellington, 2009.

[13] Nation, I.S.P., and Webb, S., Researching and analyzing vocabulary. Boston, MA: Heinle, 2011.

[14] Jean-Pierre J., R., Does size matter? The relationship between vocabulary breadth and depth, Sophia

International Review, 33, 107-120, 2011.

[15] Nation, I.S.P., and Webb, S., Researching and analyzing vocabulary, Boston, MA: Heinle, 2011.

[16] Sokmen, A. J. Current trends in teaching second language vocabulzary, in N. Schmitt and M. McCarthy (Eds.), Vocabulary: Description, Acquisition and Pedagogy, Cambridge: Cambridge University Press, pp. 237-257, 1997.

[17] Available for download at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed on 2012/03/13.

[18] For more information see: http://en.wikipedia.org/wiki/Arpabet. Accessed on 2012/03/13.

[19] Available for download at:http://www.csse.monash.edu.au/~jwb/edict.html. Accessed on 2012/03/13.

[20] Available on request.

[21] Detailed information on the CELESE Program can be found at http://www.celese.sci.waseda.ac.jp/. Accessed on 2012/03/13.