Identifying the Minimum Vocabulary Size for Academic Reading

Kazumi AIZAWA
Tokyo Denki University

Tatsuo ISO
Reitaku University

Abstract

This study investigated the relationship between learners’ vocabulary size and reading comprehension of academic texts. A reading comprehension test based on the TOEFL and an online receptive vocabulary size test were administered to 175 pre-intermediate Japanese EFL university students. The findings showed that (1) TOEFL texts could actually be considered academic, (2) 5000 to 5500 words of the JACET 8000 basic words covered 95% of the words used in the TOEFL reading section, (3) the estimated vocabulary size of 6500 is needed to score more than 60% on the reading comprehension section of the TOEFL, and (4) there were no strong relationships between vocabulary knowledge and reading sub-skills. This paper argues that the relationship between learners’ vocabulary size and reading success is not so straightforward.

1. Introduction

It is an undeniable fact that the knowledge of vocabulary plays an important role in reading comprehension. In fact, there are cases where intentional learning of vocabulary through lists and books of selected vocabulary is encouraged under a yet-to-be-proved promise of fostering better reading comprehension. The question that still remains unanswered is how large a learner’s vocabulary needs to be so that s/he would understand a text to a satisfying degree. More precisely, we do not know how many words of which frequency-level bands would enable learners to comprehend a given text to what degree. Further, the relationship between learners’ vocabulary size and reading sub-skills is also unclear. In the hope of providing beneficial information for vocabulary learning and classroom instructions, this study investigated the relationship between...
learners' vocabulary size and the reading comprehension success.

2. Overview of the previous studies

The question of how much vocabulary L2 learners of English need to acquire in order to comprehend a given text has been repeatedly asked for a few decades. In Laufer's (1992) study, based upon the results of a reading comprehension test and two sets of vocabulary tests, the threshold level hypothesis was postulated. In this hypothesis, they hypothesize the minimum amount of word knowledge to understand a given text to be 3000 word families (WF). Many researchers have supported this hypothesis, since 3000WF is a tangible and an attainable goal for ESL/EFL learners.

Nation (1993), however, counter-argues by mentioning that the hypothesis is exclusive of factors such as the learners' L1 reading comprehension skills and the background knowledge of the content, both of which are considered to affect successfulness of reading comprehension in L2. Furthermore, in Laufer (1992), the reading text and the comprehension questions to measure the learners' reading proficiency are not disclosed. This left no room for other researchers and practitioners to verify this hypothesis.

Despite the counter argument made earlier, Nation and Waring (1997) partly supported Laufer's hypothesis. By investigating this relationship between learners' vocabulary size and reading comprehension of academic texts, they discovered that 95% of lexical coverage of academic texts required 3000WF or 5000 lemmatized words.

Narrowing the focus to a more specific task, Aizawa, Mochizuki, and Yamauchi (2001) investigated the existence of a threshold level of vocabulary size in reading comprehension of the TOEFL. Their findings showed that a learner needed to have acquired vocabulary of at least 3000WF in order to correctly answer at least 60% of the reading comprehension questions. Simply stated, their study provided another support for Laufer's threshold level hypothesis. Though, unfortunately, depending on how the vocabulary size was estimated, their claim did not withstand.

Another support, albeit slightly weaker, was found in Henriksen, Albrechtsen and Haastrup (2004). By administering two types of vocabulary tests and reading comprehension tests, they revealed that there was a certain range of vocabulary sizes that did not show strong relationships with reading comprehension. It also showed that a vocabulary size outside of the range functioned as an indicator of how well, or how poorly for that matter, a text would be understood.

Although there is positive evidence for the threshold level of vocabulary size, some researchers are still skeptical about the nature of the claim, i.e., the larger the vocabulary size is, the better a learner comprehends a text. As Schmitt and McCarthy (1997) point out, the relationship between vocabulary size and reading comprehension is highly complex, since many factors are intricately intertwined and vocabulary is one of them. Further, Urquhart and Weir
(1998) denies the existence of a general threshold level, since the amount of vocabulary required for an understanding of a given text would depend on reading tasks and the proficiency levels of learners.

Taken together, as Nation (2001) summarizes, the threshold level hypothesis can be interpreted in two ways: (1) It can be viewed as a solid border line above which vocabulary knowledge ensures successful reading comprehension; and (2) It is merely a guideline which suggests the possibility of successful comprehension of a text. All in all, it is not yet clear in what way a learner's vocabulary size affects comprehension of a given text.

Theoretically, the relationship between the vocabulary and the reading comprehension has been accounted for in three ways as seen in Nagy (2005), Koda (2005), and Nation (1993). In the first of these accounts, characterized with an instrumental point of view, knowledge of vocabulary is seen as a fundamental factor that enables understanding of a text. It presupposes that acquiring larger vocabulary leads to better comprehension of a text.

The second account of the relationship views vocabulary as one of the background knowledge called for a better reading comprehension. In this view, the knowledge of vocabulary, among others, facilitates better comprehension of a written text. According to this definition, it can be said that vocabulary knowledge is important, but it cannot single-handedly facilitate successful reading comprehension, as it is merely part of the background knowledge.

Lastly, the relationship is accounted for by viewing the knowledge of vocabulary and the successful reading comprehension as separate manifestations of brain's functions. To put it differently, both vocabulary knowledge and successful reading are two of the language aptitudes, and they are interrelated but knowing more words does not necessarily cause better understanding of a text.

As has been seen so far, the relationship between learners' vocabulary size and the successfulness of reading comprehension is not as straightforward as it looks. However, for the purpose of simplicity and ease, we temporarily adopted the first of the three theoretical definitions above. Using this hypothesis allowed us to directly investigate the relationship between learners' vocabulary size and the success in reading comprehension of academic texts. We frankly admit that this presupposition is controversial, but we believe this is the only way to verify the threshold level hypothesis.

3. The Study

3.1 Purpose

This study, limited to academic texts, was designed to investigate the relationship between learners' vocabulary knowledge and their performance in reading comprehension test of academic texts. For this aim, the following three research questions were asked.

(1) What frequency levels of vocabulary are used in a TOEFL reading section?
How large vocabulary is needed in order to score 80 percent on a TOEFL reading section?

(3) What kind of relationship can be observed between the sub-skills of reading comprehension and the vocabulary profiles of frequency bands?

Having mentioned that the focus is on academic texts in this study, the first question was asked as to determine the texts from TOEFL actually fall into the academic category.

3.2 Subjects
A total of 175 Japanese university students participated in the study. Ninety-three of the subjects were majoring in English at one university, and 82 were from another university where they majored in engineering.

3.3 Materials
Reading comprehension test: The test comprised of three reading texts with comprehension questions from a reading section of TOEFL. The test was identical to the one used in Aizawa, et al. (2001). The basic statistics of the test is shown in Table 1. The Flesch-Kincaid readability scores indicated that the texts were suitable for American high school freshmen. After eliminating vocabulary-related questions, there remained 7 or 8 questions for each text. A correct answer to a question was granted 1 point, and the total score was 23.

Vocabulary test: As a measurement of vocabulary knowledge, flashTDUVLT (Aizawa & Iso, 2007) was used. This is a web-based receptive vocabulary test that estimates the size of overall vocabulary up until 8000 words and the frequency-level bands of learners. Thirty randomly chosen vocabulary items from each of the eight levels of JACET 8000 were used in the test, totaling 240 test items. Three test items of the same level are presented together with multiple choices of six Japanese words, and test takers are to match the three words with their Japanese equivalents out of six choices. The maximum score of 240 is the equivalent of the estimated value of 8000 words.

Table 1. Statistical analysis of reading texts

<table>
<thead>
<tr>
<th>Topic</th>
<th>Word Count</th>
<th>N of Qs</th>
<th>Flesch-Kincaid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text A Agricultural Revolution</td>
<td>201</td>
<td>8</td>
<td>11.0</td>
</tr>
<tr>
<td>Text B Migratory Birds</td>
<td>235</td>
<td>8</td>
<td>6.9</td>
</tr>
<tr>
<td>Text C Glacier National Park</td>
<td>229</td>
<td>7</td>
<td>9.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>669</strong></td>
<td><strong>23</strong></td>
<td><strong>8.9</strong></td>
</tr>
</tbody>
</table>

3.4 Method
For the research question 1, the reading texts were analyzed with BNC Range so that they can be compared with the results from Nation’s (2004) analysis of academic texts. The tests were further analyzed using v8an (Shimizu, 2004) to discover what proportion of the vocabulary used
in the reading texts can be covered by each of the eight levels of JACET 8000.

In terms of research question 2 which asked the relationship between the vocabulary size and an overall reading comprehension test score, the vocabulary test and the reading test were administered on two different occasions.

Lastly, for the third research question, the reading comprehension questions were divided into groups of sub-skills required to correctly answer each question. Then the scores of each sub-skill and the results of the vocabulary test were compared by the sub-skills.

3.5 Results

The first research question dealt with the frequency levels of vocabulary used in a TOEFL reading section. Figure 1 shows how each level of the British National Corpus (BNC) covers the texts from the TOEFL reading section and the academic text corpus compiled by Nation (2004). Even though the Nation’s analysis was from 1000 to 3000 BNC levels, the figure seems to show a very similar tendency in terms of the coverage rate. From the results, it is assumed that the vocabulary used in a reading comprehension section of the TOEFL does not differ from the academic text corpus when compared in view of the BNC levels’ coverage rates.

In Figure 2, the result of another frequency-level analysis based on JACET 8000 is displayed. Looking at the average of the three texts, the figure shows that 3000 words in JACET 8000 covered almost 90% of the words used in the texts, and 5000 words did 95%. Figure 2 also shows the results of the frequency-level analysis for each independent text. Even though there are slight differences among the texts, the shift of the coverage rates through 1000 to 8000 levels for each text were very similar. Moreover, when compared with the overall text coverage, any of the three texts did not show strong deviations from the average.

In sum, it was found that the texts from the reading section of the TOEFL were very similar to the academic corpus of Nation (2004) in terms of the frequency-level coverage. It was also
discovered that the frequency-level coverage of the three texts was similar to each other and learners need to know 5000 to 5500 words of JACET 8000 to cover 95% of the vocabulary used in the TOEFL reading comprehension questions.

The second question asked how large of a vocabulary is needed to correctly answer 80% of the TOEFL reading comprehension questions. Before analyzing the results of the vocabulary test and the reading comprehension test, 23 subjects were eliminated from the data as they failed to participate in either of the measurements. Further, since the reading comprehension test involved multiple-choice questions with 4 alternatives, randomly answering all the 23 questions would have yielded a score of 5.75. To avoid such data tampering the results, two subjects who scored less then 6 out of 23 were also eliminated from the following analyses.

Table 2 shows the descriptive statistics of the two measurements. For the reading comprehension test, there were no subjects who scored the possible maximum of 23 points. Taken together with the mean score of 13.6, it can be said that there was no ceiling effect for the test.

To answer research question 2, the subjects were first divided into three groups depending on the percentage of correct answers; 80% and above, 60 to 79%, and below 60%. Then within each group, the subjects were further divided into the estimated vocabulary size groups of below 6000, 6000 to 6499, and 6500 and above. The distributions of the subjects as well as the result of chi-squared analysis are shown in the left side of Table 3. When focusing on the subjects who scored more than 80% in the reading comprehension test, the number of the subjects who were

<table>
<thead>
<tr>
<th>Est. Vocabulary Size (words)</th>
<th>N 6000- 6499 6500+</th>
</tr>
</thead>
<tbody>
<tr>
<td>80+</td>
<td>14 1 △ 2 11 ▲</td>
</tr>
<tr>
<td>60-79</td>
<td>58 14 △ 22 22</td>
</tr>
<tr>
<td>60-</td>
<td>78 36 ▲ 25 17 ▽</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correct Answers (%)</th>
<th>N 6000- 6499 6500+</th>
</tr>
</thead>
<tbody>
<tr>
<td>60+</td>
<td>72 15 △ 24 33 ▲</td>
</tr>
<tr>
<td>60-</td>
<td>72 36 ▲ 25 17 ▽</td>
</tr>
</tbody>
</table>

△: Significantly larger than the expected value.
▽: Significantly smaller than the expected value.
estimated to know more than 6500 words was 11. The chi-squared analysis showed that the number was significantly larger than the expected value of 4.66 (14 subjects equally divided into 3 groups), whereas there were only three subjects with less than 6500-word knowledge. The result seems to suggest the existence of a threshold level between 6000 and 6500 words. Moreover, it was apparent that there were 39 subjects (22 and 17) with more than 6500 words whose reading comprehension test scores were below 80%. What this indicates is that the vocabulary knowledge of 6500 and more is necessary to score 80% in the TOEFL reading comprehension section, although knowing that much vocabulary does not necessarily ensure that a learner would score more than 80%.

Since the number of the subjects who scored more than 80% in the reading comprehension test was rather small compared to the lower-criterion groups, the subjects in the 80% and 60 - 79% groups were merged to form a new group. The distribution of the subjects is presented in the right side of Table 3. As can be seen, when the two groups of the subjects who knew more than 6500 words were compared in terms of the percentage of the correctly answered comprehension questions, those who scored more than 60% outnumbered its counter part which did not reach the 60% in reading comprehension test (33 and 17 respectively). Shifting our focus to the subjects with less than 6000-word knowledge, it was revealed that the number of such subjects who scored more than 60% in the comprehension test was less than half, when compared with the number of subjects who scored 60% or lower. These results again led us to suggest that there may be a threshold level of 6500 words even when the criterion of the reading comprehension score was lowered from 80% to 60%.

Taken together, the results of the vocabulary test and the reading comprehension tests demonstrated that the estimated vocabulary size of 6500 words is needed to score more than 60% in the TOEFL reading comprehension section, although knowing 6500 words does not necessarily mean that reading texts will be understood at a satisfactory level.

<table>
<thead>
<tr>
<th>Sub-skills</th>
<th>No. of Qs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohesion</td>
<td>5</td>
</tr>
<tr>
<td>Inferencing</td>
<td>3</td>
</tr>
<tr>
<td>Getting local information</td>
<td>2</td>
</tr>
<tr>
<td>Getting global/local information</td>
<td>3</td>
</tr>
<tr>
<td>Getting the main topic/general idea</td>
<td>3</td>
</tr>
<tr>
<td>Scanning</td>
<td>4</td>
</tr>
<tr>
<td>Skimming</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 4. Percentages of correct answers by vocabulary size
To answer the third research question, it was necessary to decide what sub-skills were called for when answering the 23 reading comprehension questions. Five English teachers were asked to take the reading comprehension test and then to make judgments as to what kind of sub-skills are required to answer each question. The result list of the seven sub-skills is shown in Table 4. For each sub-skill, the number of correct answers was calculated by three vocabulary size groups: less than 6000, 6000 to 6499, and over 6500 words. The comparisons of the groups for the seven sub-skills are shown in Figure 4. Though cohesion was expected to be a difficult skill in reading comprehension, the result showed that the questions that required this skill were found to be easiest in all three groups of the subjects. On the other hand, the most difficult questions were the ones that called for inferencing skills where only less than 50% of the questions of this type were correctly answered by all the three groups.

The sub-skill that seemed to be most closely related to the vocabulary size was ‘getting global/local information’ where the differences between the vocabulary size groups were the largest. This means that when the subjects were to find specific information somewhere in the body of the text, the knowledge of larger vocabulary helped them to find what they were looking for. Although not as apparent as the ‘global/local information’, questions that required ‘skimming’ also seemed to be affected by the size of vocabulary as shown in Figure 4.

4. Discussion

For the research question, this study dealt with the concept of academic text. With the daily expansion of academia, it has not always been clear as to what comprises of a list of academic words. In this study, three reading comprehension texts from the TOEFL were compared with Nation’s (2004) academic text corpus, and it was found that the frequency-level coverage of the three texts was considered to be the same with Nation’s corpus. It was also shown that to cover 95% of the vocabulary in the reading texts of TOEFL, a minimum of 5000 words in JACET 8000 is needed.

The results need to be interpreted with caution, since they do not automatically mean that having an estimated vocabulary size of 5000 words is enough to know 95% of the words in academic texts. This is partly because of the nature of vocabulary size tests. In most of the vocabulary size tests that try to estimate learners’ vocabulary size of different frequency levels, it is almost always impossible to test the knowledge of all the words of, say, 8000. Due to this fact, even though the estimated vocabulary size was calculated as 5000 words, there are still many words that may not have been acquired. In fact, this study witnessed such cases where the subjects with the estimated vocabulary size of 5000 did not know all the 30 tested words in higher frequency-level bands. Therefore, it is safe to aim for acquiring 7000 to 8000 words to attain the actual mastery level of 5000 words.

As for the second research question, the minimum vocabulary size to satisfingly
understand texts from the TOEFL reading section was suggested to be about 6500 words. Admitting the fact that JACET 8000 is a list of lemmatized words, the figure seems to be rather high compared to Laufer’s 3000WF. Two possible interpretations of this result follow.

Firstly, the result can be interpreted to support Urquhart and Weir’s (1998) claim that threshold level is dependent on tasks and subjects. Although the same reading tests were used in Aizawa, et al. (2001), the results of the current study showed a very different picture of the relationship between vocabulary size and the reading comprehension. Since both the vocabulary test and the subjects are different, the results cannot be directly compared. Even so, the findings of the two studies seem to support Urquhart and Weir’s claim.

Another interpretation is that the results of the current study support the claim by Nagy (2005). The results observed in this study did not clarify the direct relationship, not to mention the causality, between vocabulary size and reading comprehension. Even so, there were strong tendencies that larger vocabulary size may facilitate better understanding of texts. It was also evident that vocabulary is not the single most affective factor in successful reading comprehension. Further research on the other factors as well as the relationships between such factors is needed.

Unfortunately, we could not observe strong relationships between vocabulary size and reading comprehension sub-skills. Although the findings showed that questions that required inferencing skills were difficult regardless of a learner’s vocabulary size and that vocabulary size did affect the success of gathering information globally and/or locally to answer a given question, those findings did not yield a whole picture of the relationship. One of the possible explanations for this is that the questions accompanied the reading texts were too difficult for the subjects. In fact, there were no questions that majority of the subjects correctly answered.

Another explanation is that the types of sub-skills may not be the strong determinant of the successfullness of comprehension questions. In other words, the difficulty of a given comprehension question may derive from the interaction of many factors, such as vocabulary, parts of the text from which a question is based on, and the inherent difficulty of the skill types. For now, we may have to rely on a test maker’s intuition when estimating the difficult of reading comprehension questions.

5. Implications for future research

As stated in the beginning of this study, the relationship between learners’ vocabulary size and success of reading comprehension tasks is not straightforward. There seems to be other additional factors to be considered to explain reading comprehension. For further studies, we should consider other factors like learner’s grammar knowledge, automaticity, background knowledge, and so on.

Another limitation is that only three academic texts and one set of vocabulary tests were used in this study. In order to extrapolate the results, more studies with other academic texts
from various fields are needed. Also, different vocabulary tests based on different lists of vocabulary and other methods of measuring learners’ vocabulary size may also reveal useful information to clarify how important a role vocabulary plays in reading comprehension.

References