The Interrelationship Among Word Frequency, Learner Behavior in a Vocabulary Size Test, and Teachers' Perception of Difficult Words

Tatsuo ISO
Reitaku University
Kazumi AIZAWA
Tokyo Denki University

Abstract

This study investigated the relationship among word frequency, learner performance, and teacher intuition about difficult words. It was expected that gaining deeper insights into such relationship would demonstrate the frequency of words reflected on the difficulty of them. Firstly, by employing a vocabulary size test that takes learner confidence into account, 180 university learners’ vocabulary size was measured. The results of the test confirmed that the scores corrected by the degree of confidence were more sensitive to frequency levels than the raw scores. Secondly, when the distribution of difficult words judged by teachers were investigated across frequency levels, the number of difficult words were in accordance with the frequency levels. Lastly, the relationship between learner performance in vocabulary size test and teacher intuition about word difficulty was investigated. As a result, a strong correlation was revealed. The result is interpreted to show that the teachers were in fact capable of predicting difficult words. Combining all the results together, this study confirmed a close relationship between the frequency of words and their difficulty.

1. Introduction

Why are certain words difficult to learn? This is a question that has been asked many times. Sometimes it is raised within learners themselves. At other times, learners pose such questions to language teachers. Ever since researchers became focused on the importance of vocabulary knowledge (Meara, 1980; Cody & Huckin, 1997), there have been attempts to find the answer to the question. Laufer (1997), for example, summarized the results of earlier studies and listed theoretical “intralexical” factors that were thought to be closely related to word difficulty. To put it differently, the discussion of word difficulty sought its answer within a word itself.

With the development of large-scale corpora like BNC and vocabulary size tests that use
corpus data as item pools, it is now possible to estimate learners’ vocabulary size. Furthermore, since the corpora provide information about the frequency of each word, vocabulary size tests are also capable of estimating the size of vocabulary that belongs to an arbitrary frequency band. On the surface, the results of such vocabulary size tests seem to show how hard it is to learn words within a certain frequency band. It is even almost natural to interpret the results in this manner. After all, not knowing a word could be a result of the word being difficult to learn. One may still argue, though, that the frequency level of a word is not necessarily related to its level of difficulty. This argument is based on the possibility that learners’ poor performance could also be a result of not having seen the word before. Although several studies were conducted to clarify the relationship between the frequency level of words and their difficulty, it still is not entirely clear how the frequency of a word is related to difficulty, which in turn supports or refutes such an argument. Hence, this study was conducted to gain deeper insights into the relationship between frequency and difficulty of words.

2. Literature Review

A few studies have attempted to predict the difficulty of words with frequency levels. The general agreement seems to be that the difficulty of words is closely related to the level of frequency (e.g., Aizawa, 2006; Aizawa & Iso, 2007). In these studies, the researchers tested Japanese university students’ knowledge of English words of 1000 to 8000 frequency levels using vocabulary tests based on JACET 8000 (hereafter, J8; The Japan Association of College English Teachers, 2003). The findings showed that the number of words known to the learners steadily decreased up to 4000 or 5000 frequency level, but such a declining pattern of knowledge was not observed among the lower frequency bands. Such findings led the researchers to conclude that a word’s frequency level could partially explain the difficulty of the word.

Okamoto (2008) reported a similar tendency. One of the most prominent differences from the earlier studies was that words of even lower frequency were included in the vocabulary size test. The lowest frequency in the study was 15000 level. Here again, the conclusion was that the words of lower frequency were difficult.

There are two problems in the above studies. One of them is an underlying assumption that the number of words known to learners is analogous to the degree of ease in learning a word of a certain frequency level. In the previous studies, learners’ poor knowledge of words at a certain level of frequency was interpreted to indicate that the degree of difficulty of the words was high. One may argue that such an assumption is false. In such case where learners’ vocabulary size test results are poor for lower frequency words, it could also mean that the learners may not even have opportunities to learn these words because they are low in frequency.

The other is concerned with the guessing in the tests. In the previous studies, the materials used to measure the subjects’ vocabulary size were receptive, multiple-choice vocabulary tests.
As is indicated by Iso and Aizawa (2008), such tests are susceptible to guessing. If the results of the vocabulary size tests were affected by such test-taking skill, the relationship between the frequency of words and difficulty of them might well be obscured.

In terms of the first problem, an encouraging study of this situation is found in Aizawa and Iso (2007) where Japanese- and English-speaking teachers of English were asked to make a dichotomous (easy/difficult) judgment for 480 English words. The instructions given to the teachers were to divide the words into two groups in terms of difficulty. Although there was an exception in the data, the general tendency was that the lower the frequency of the words became, the larger the number of difficult words was. The results indicated that teacher intuition was an indicator of word difficulty. Unfortunately, the teachers' intuitive prediction about difficult words was not directly compared with the learners' performance in the vocabulary size test conducted in the study.

In order to clarify the relationship between the frequency of words and their difficulty, then, the following issues have to be dealt with. Firstly, the guessing factor in the vocabulary size test was eliminated as much as possible. Through an original vocabulary test, Iso and Aizawa (2008) revealed that the confidence in form-meaning matching task decreased in relation to frequency levels, whereas the number of correct response did so only partially. Thus, it is quite possible that without measuring the degree of confidence, vocabulary test results may be unreliable. In this study, therefore, both raw scores and confidence-based corrected scores were considered as results of vocabulary size test.

Secondly, it is necessary to make the word difficulty judged by teachers directly comparable with the results of vocabulary size test taken by learners. Therefore, when collecting data on teacher intuition, the authors set up the questionnaire so that the teachers judged the word difficulty with a common standard. Since the learner performance in a vocabulary size test was to be compared, a difficult word in this study refers to a word that is difficult for learners, but not for the teachers themselves or for them to teach.

By dealing with these issues, this study aims to clarify the rather obscure relationship between word frequency and its difficulty.

3. The Study

3.1 Purpose

As is apparent from the literature review, further investigation into the relationship between frequency level and word difficulty is in order. Specifically, this study attempts to clarify the intricate relationship among word frequency, learner performance in vocabulary size test, and teacher’s intuition about difficult words. This was done by directly comparing the results of vocabulary size test administered to learners and those of a vocabulary difficulty questionnaire completed by teachers.
For the aforementioned purpose of this study, the following research questions were asked.

1. How do the corrected scores of a vocabulary test reflect the frequency levels of target words and how do they differ from the raw scores of the same test?
2. How does the number of difficult words judged by teachers change in relation to J8 frequency levels?
3. How well can language teachers predict the difficulty level of words for learners?

3.2 Participants

There were two groups of participants who were assigned different tasks. One group of participants consisted of 180 university students from two different institutions. About one-third of the participants were engineering majors, while the remaining participants were foreign language majors. All of the participants in this group (hereafter learners) took two forms of a vocabulary size test over a period of two weeks.

The participants in the other group were teachers of English at various levels and institutions. Within this group, 27 Japanese teachers of English formed a subgroup and 14 native speakers of English formed another, totaling 41 participants (hereafter teachers). They all agreed to participate in and completed an online survey concerning word difficulty.

3.3 Materials

Vocabulary size test. For learners, two forms of FlashVLT (random version) were administered. In this test, the learners were to match the three English words with definitions from a list of six Japanese words displayed on a computer screen at one time. Unlike other widely-known vocabulary size tests, FlashVLT requires test-takers to mark their confidence at one of three levels upon answering. The three levels include confident, not so confident, and not confident at all. (For more details, see Iso and Aizawa, 2008). By using this test, it was possible to correct scores to minimize the number of correct responses by guessing.

By completing the two forms, they were tested for receptive knowledge of 480 words (60 words x 8 levels) in total. The test items were randomly displayed on the computer screen irrespective of the original levels in J8.

Another vocabulary test, namely Yes/No vocabulary test (Meara, 1992) was also considered, as it is known to effectively eliminate guessing by using non-words. However, as mentioned earlier, the results of this test and that of teachers’ prediction had to be directly comparable by using the same target words. Judging the difficulty of non-words was thought to be confusing for teachers. For this reason, the authors did not select the Yes/No vocabulary test as the measurement of vocabulary size for the present study.

Word difficulty questionnaire. For teacher participants, an online word difficulty questionnaire was administered (see Figure 1). In this questionnaire, the participants were shown
lists of randomly chosen words and asked to identify words that they believed would be difficult for their students to learn. The words to be judged were the same 480 words in FlashVLT that was administered to the learner participants.

After providing biographical data, including name, institutional level, and the length of teaching experience, the teachers were shown a list of 20 words that were randomly chosen from the 480 words. While going through the list, the participants were asked to click on words that they believed would be difficult for their students. The exact instruction given to the teachers was, “On the screen, you will see 20 English words at a time. Please click on the words which you think are difficult for your students. You can click on as many or as few words as you’d like.” Through repeating this process 24 times, the teachers completed the process of identifying difficult words among the 480 words.

3.4 Results

The general results of the vocabulary size test are shown in Table 1 and Figure 2. To calculate the estimated vocabulary size, the number of correct responses was divided by the total number of test items and then multiplied by the number of words contained in J8. We can assume from these results that FlashVLT was a suitable vocabulary size test for the learners in this study, even though there were a few whose vocabulary sizes were smaller than three standard deviations from the mean. Such participants could be seen as outliers; however, it was our decision to keep them to maintain variety in the data.

The first research question asked how confident-based corrected scores, as opposed to raw scores, reflected the frequency levels of target words. To answer this question, the number of correct responses for each J8 level was compared with the number of confident correct responses. Table 2 and Figure 3 show the result. As is obvious from the figure, the number of confident correct responses was smaller than

![Figure 1. Sample screen of the word difficulty questionnaire](image)

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive statistics of the vocabulary size test</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>180</td>
<td>5435.5</td>
<td>916.8</td>
<td>1967.0</td>
<td>7376.0</td>
</tr>
</tbody>
</table>
that of correct responses at each level. Further, such differences intensified at lower frequency levels, with the exception of the 8000 level. The differences between the raw scores and corrected scores were found significant as the T-test revealed \( t (179) = 9.56, 15.96, 18.50, 23.45, 24.86, 29.22, 31.17, 26.96 \), in the order of 1000 to 8000 level, \( p < .001 \). Additionally, one-way ANOVA of with-in subject design revealed that the main effect of frequency level was significant for both raw scores and corrected scores (\( F (7, 1253) = 1603, p < .01 \); \( F (7, 1253) = 1617, p < .01 \); respectively). The pair-wise comparisons showed that for raw scores, the scores of one frequency level decreased statistically significantly compared with the next level of frequency (\( p < .01 \)). The only exception was between 4000 and 5000 where the differences between the score for those levels were not significant (\( p > .01 \)). As for corrected scores, the score of each frequency level was significantly higher when compared with that of a lower frequency level (\( p < .01 \)). Yet, such difference was not observed between levels 7000 and 8000 (\( p > .01 \)). What this indicates is that the corrected scores curved by the reported confidence value was more sensitive to the frequency level of words as they maintained significant decline from one frequency level to the next up to 7000 levels. This is especially so after 4000 level as the corrected scores for such levels kept declining significantly. It is important to keep it in mind, however, that the sensitivity to the frequency level does not automatically mean the sensitivity to the difficulty of words. This is so because the difficulty of the target words remains unknown.

![Figure 2. Distribution of the observed vocabulary size of the learners](image)

<table>
<thead>
<tr>
<th>J8 level</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>7000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>56.9</td>
<td>50.7</td>
<td>45.2</td>
<td>37.5</td>
<td>37.3</td>
<td>33.8</td>
<td>32.9</td>
<td>31.8</td>
</tr>
<tr>
<td>Corrected</td>
<td>54.2</td>
<td>43.7</td>
<td>36.1</td>
<td>26.1</td>
<td>24.9</td>
<td>21.4</td>
<td>18.8</td>
<td>18.5</td>
</tr>
</tbody>
</table>

Table 2

Mean of raw scores and corrected scores

![Figure 3. Correct and confident correct responses](image)
number of words predicted to be difficult was counted for each teacher. The mean for all the J8 levels are shown in Table 3. From this table, it is apparent that the number of difficult words increased as the frequency level of the word decreased regardless of the native language of the participants. The only exception was the 4000 and 5000 level for the NS teachers, which showed a slightly higher mean of 18.1 for the 4000 level as opposed to 15.7 for the 5000 level. Even though the 480 target words were shown to the teachers randomly, the result shows that the teacher’s judgment about word difficulty was in accordance with the frequency levels within the J8.

The third research question posed was if the teachers could predict difficult words for their learners. To investigate the relationship between teacher intuition and learner performance, the correlation coefficient was calculated between the number of teachers who judged each word as difficult and the number of learners whose response to the word was correct. Since the easiest word should have zero on the teacher side and 180 on the learner side, it was expected that the correlation coefficient would be negative. In fact, the correlation coefficient was -0.81 (p < .001), which shows a strong correlation. This serves as a piece of evidence to show that the teachers could predict difficult words for their learners.

Even with the strong correlation, a closer examination of the data revealed that there were a number of cases where the teachers and the learners did not agree on word difficulty. Table 4 shows the numbers of words that were judged as difficult by more than 50% of the teachers and were not correctly answered by more than 50% of the learners. The criteria was set at 50% so that the data would be sensitive enough to individual differences among the participants but at the same time robust enough to avoid being affected too strongly.

It is evident from the table that there were no difficult words commonly identified by both the teachers and the learners among the most frequent 3000 words.

| Table 3 |
| Number of difficult words predicted by teachers |
| J8 level |
| 1000 | 2000 | 3000 | 4000 | 5000 | 6000 | 7000 | 8000 |
| NNS | 2.3 | 9.2 | 15.9 | 25.1 | 27.3 | 34.9 | 37.1 | 38.9 |
| NS | 0.6 | 5.4 | 6.1 | 18.1 | 15.7 | 21.6 | 24.2 | 28.9 |
| Total | 1.8 | 7.9 | 12.5 | 22.7 | 23.4 | 30.3 | 32.7 | 35.4 |

The answer to the second research question will disclose such information.

The second research question asked how the number of difficult words judged by teachers changed in relation to J8 frequency levels. To answer this question, the
The general tendency was that the teachers overestimated learners’ knowledge of words in higher frequencies of 2000 to 5000 levels, and underestimated in lower frequencies. The differences between the two groups indicated such tendency with the values being negative and positive, respectively.

Another tendency was found in the number of difficult words commonly identified by both teachers and learners. Although there were no such words among 1000 to 3000 levels, the number increased towards the lowest frequency levels. In other words, even though it seems easier to predict difficult words among less frequent words, teachers need to pay extra care when dealing with high-frequency words because teachers may tend to overestimate learners’ vocabulary knowledge for such words, and teacher expectation may not be met.

To sum up, the teachers do have ideas about a given word and its level of difficulty for the learners, and they are, for the most part, quite accurate. Even so, the chances of them being accurate in their estimations decrease as the frequency level of vocabulary increases.

4. Discussion

The current study set out to investigate the relationship among frequency of words and their difficulty. Firstly, a comparison was made between the raw scores of the FlashVLT and the scores corrected with the reported confidence. In terms of the number of correct responses with confidence, the results were in accordance with those of Iso and Aizawa (2008), where smaller number of words was tested in high-to-low frequency order. With increased item number and random question sequence, this study confirmed that the number of confident correct responses decreased as the words’ frequency became lower. Moreover, this study revealed that the raw scores were statistically significantly higher than the scores corrected with the confidence value at each level of J8.

Interpreting these results in terms of the relationship between the frequency level and the word difficulty, the corrected scores for each J8 level were found to be more susceptible to the frequency level. Additionally, since the observed differences between raw scores and corrected scores were significant at all levels of frequency, we argue that the vocabulary size test in this study succeeded in controlling one aspect of learner behavior during a vocabulary size test by eliminating guessing factor to satisfactory level.

As for the teacher perception of word difficulty, the survey results were in accordance with the word frequency: the teachers judged words of higher frequency as easy and those of lower frequency as difficult. It is important here to stress that the teachers were asked only to identify difficult words for a specific group of learners, and the results were in accordance with word frequency levels. Interestingly, it has been reported that even professional linguists who natively speak English fails to distinguish the frequency level of English words (Alderson, 2007). If this is true, we can assume that the difficulty judgment made by the teachers in this study was based on
the difficulty of words, but not the frequency level. In other words, the difficulty of words judged by the teachers is closely related to the frequency level of these words.

These findings in this study seem to indicate a close relationship between word frequency and difficulty. What ties these findings together was the strong correlation between the learner performance and the difficulty judgment by the teachers. This completes the seemingly triangular relationship among the frequency level, learner performance, and the word difficulty: the learners’ poor knowledge of lower frequency words reflects those words’ difficulty, as those words were judged difficult by the teachers who made such judgment based the word difficulty, not the word’s frequency.

Upon closer look, however, it was found that there were 66 words that were identified on opposite sides as difficult. Of these words, 37 words were from the teacher side and 29 from the learner side. Although it would be informative to analyze the reasons behind such differences, we refrain from any speculation toward this phenomenon in this study. This is because of the method with which the data was collected. In this study, the learners’ data was collected by a multiple-choice vocabulary size test. The teachers, on the other hand, were not informed of the format of the vocabulary size test that the learners took. If the teachers were aware of the test format, it would have affected the outcome of the survey as the test measured only the learners’ receptive vocabulary knowledge, which is the weakest form of vocabulary knowledge (Laufer & Goldstein, 2004). This issue needs to be taken into consideration when a study of similar interest is carried out in the future.

Despite the results, it is not our intention to claim that frequency equals difficulty. For one, the causality has yet to be found between the two. In addition, their relationship might be dependent on a learner’s proficiency level. It is not very difficult to imagine a situation where a beginning level learner of English finds it difficult to learn enough. Until s/he becomes familiar with the spelling-pronunciation correspondences, such word remains as a somewhat strange word. According to J8, the frequency order of enough is 250. Apparently, the concept of word difficulty is far more complex and cannot be determined by mere frequency level. Therefore, the findings of this study are by no means generalizable. Even so, this study revealed that the frequency level of a word is closely related to the difficulty of a word, and for certain level of learners, the information about word frequency can function as a strong indicator of word difficulty.

Acknowledgment

We would like to show our gratitude to all the teachers who provided us with the valuable data and constructive criticisms for the word difficulty questionnaire. We are also indebted to Dr. Tess Fitzpatrick and the three anonymous reviewers who provided us with comments and suggestions.
References


Aizawa, K., & Iso, T. (2007). Estimating word difficulty: The divergence from frequency levels. ARELE, 18, 111-120.


