Estimating Word Difficulty:
The Divergence From Frequency Levels

Kazumi AIZAWA
Tokyo Denki University

Tatsuo ISO
Reitaku University

Abstract

This study attempts to clarify the difference between word frequency and word difficulty for Japanese EFL learners. The first experiment is designed to investigate learners' vocabulary knowledge at each frequency-band based on JACET 8000. The subjects took a bilingual computerized vocabulary size test and their scores were analyzed according to the frequency bands from Level 1 to Level 8, each level containing 1000 words. The results show that frequency bands can provide an implicational scale for the difficulty of words up to Level 4. Beyond Level 5, however, the scores for each level remained the same. The second experiment asked 40 English teachers about the difficulty of the target words used in the first study. The results showed that the number of the words judged as "difficult" by more than three of the teachers increased gradually up to Level 4, but beyond this level the number of difficult words in each level remained the same. However, when we took a different data analysis approach by modifying the definition of "difficulty" to "judged as such by more than 50% of the teachers", we discovered a conflicting result; the "difficult" words continued to increase even after Level 4. Conclusions are drawn about these two apparently contradictory results.

1. Introduction

Recently an increasing emphasis has been put on learners' vocabulary knowledge (e.g. Nation, 2001). Some studies have defined vocabulary knowledge as an important part of language competence, and investigated the relationships between vocabulary knowledge and the development of English proficiency (Katagiri, 2001). There has been an implicit assumption among practitioners that the more words learners know in a second language, the more proficient
they are in that language. Therefore, a basic question in the diagnosis of a learner’s proficiency is how many words the learner knows. It is no wonder that practitioners support this assumption, as it is entirely plausible to them in the light of their teaching experiences.

Furthermore, there is tacit consent among practitioners that the difficulty of a word increases in inverse proportion to its frequency. In other words, the more frequently learners encounter a word, the easier it is for them to acquire (e.g. Read, 1988).

Mochizuki, Aizawa & Tono (2003), through their observations as practitioners, liken the relationship between learners’ vocabulary knowledge and word frequency to a pyramid with missing blocks. By doing so, they illustrated their assumption that even among numerous high-frequency words, there remain some unknown words. However, there is a handful of infrequent words that they have acquired for various reasons.

The assumption of Mochizuki et al. is plausible, but has not been empirically proven yet. This study, therefore, is designed to investigate the relationship between frequency of words and their difficulty, and between word frequency and perceived word difficulty by English teachers.

2. Overview of previous studies

Attention to the frequency of words seems to be crucial with second language learners, especially those in EFL countries such as Japan with little or no exposure to communicative needs. However, surprisingly few studies have focused upon the relationship between frequency count and difficulty of words. Aizawa (2006) investigated the scores of a vocabulary test based upon the frequency bands of the JACET 8000 list. The results showed that up to the 4000-word level the mean scores of the tests declined in proportion with the frequency-band, but were unpredictable beyond this, making it impossible to describe these words as “increasingly difficult”. These results led to the conclusion that a threshold may exist at around the 4000 word level: below this level learners need to know words as core vocabulary, while beyond this level word frequency is of less importance. One potential weakness in the study was that the homogeneity of the subjects. If the subjects in the study had been more proficient, a threshold level might have been discovered at a higher frequency level.

Aizawa’s conclusions are further called into question by Fujii’s (2006) contradictory findings. Using a survey of word difficulty, Fujii asked the subjects to rate the difficulty of given words. These values of difficulty were adjusted according to their English proficiency, in such a way that the rates provided by higher proficiency learners weighed more than the rates of monitors with lower proficiency. He concluded that it was not possible to estimate the difficulty of the word by its frequency of occurrence. However, the validity of this weighing system is questionable, and his findings could be skewed as a result.

From a slightly different angle, Mizumoto (2006) investigated the relationship between word frequency and familiarity by using his own receptive vocabulary test based upon JACET 8000.
In this test, he randomly chose 30 words from each of the eight levels (240 words in total) and translated them into Japanese. The subjects chose one of the four English words for each test item. Rather than tallying each subject’s scores, their answers to each test item were seen as “familiarity.” His data showed two interesting findings: (1) the frequency band can be categorized into four groups: Level 1, Level 2 and Level 3, Level 4 and Level 5, and beyond Level 6, and (2) word frequency did not correspond with learners’ familiarity with the words. However, his test seemed to have some weaknesses, as it seemed to measure a productive aspect of vocabulary rather than receptive, and the number of the test items was limited to 30 items for each level.

In an attempt to move forward from the recent studies described above, this study is designed to overcome three hurdles through (1) obtaining subjects with different levels of proficiency, (2) adopting a vocabulary test which measures receptive aspects of vocabulary knowledge with more test items for reliability, and (3) adopting plausible methods of analyzing the data obtained from the subjects. With these three improvements, this study aims to further investigate the relationship between the frequency of words and their difficulty.

For operational reasons, “difficulty” is defined as “the difficulty index obtained from a vocabulary test” in this study. Some may claim there are many aspects of vocabulary knowledge and the difficulty of words cannot be estimated by a single test. However, what is important here is a rough estimation of learners’ vocabulary size, so this claim is beyond the scope of this study.

3. Study 1

3.1 Purpose

This study was designed to measure learners’ word knowledge of each frequency band by using JACET 8000 as a database source.

3.2 Subjects

The subjects were 164 Japanese students studying at two different universities (142 engineering majors and 22 English majors).

3.3 Method

To measure learners’ vocabulary size, the subjects were asked to complete two sets of Tokyo Denki University Vocabulary Levels Test (TDU VLT) online (Aizawa & Iso, 2004) on two different occasions (See Appendix 1). This web-based test is designed to assess the breadth of receptive vocabulary knowledge and adopts the format of the VLT (Schmitt, Schmitt, & Clapham, 2001). Three sets are currently available, and each set consists of eight different levels that mirror Level 1 to Level 8 of JACET 8000. At each level, there are 10 clusters of three Japanese and six English words, totaling 30 definitions and 60 words. Test takers have to make three pairs of Japanese-English equivalents by choosing three English words out of six from each cluster.
There is a time limit of 3 minutes per level, and it takes at most 24 minutes to complete a set. In this study, two sets of the tests were administered and the total raw score of each level was 60, which made the overall possible maximum of 480 (60 points by 8 levels).

3.4 Results
The scores learners achieved at each level of the two sets of the tests were added together, divided by the possible maximum score of 480, and then multiplied by 8000 to calculate their estimated vocabulary size. Although their actual vocabulary size might be larger because of words known outside of the JACET List, our focus was on the subjects' profile in terms of the JACET frequency bands.

The overall mean vocabulary size out of the 8000 words was 4473.5 words (SD=944.4). Figure 1 shows the mean of each level, while the vertical lines show the SDs. As we can see from this figure, the mean was the highest at Level 1 (54.4) and decreased steadily toward Level 4 (28.0). However, it leveled off from Level 4 to Level 5 (27.9) with a deviation of only minus 0.1. It again decreased at Level 6 (25.1), but at Level 7 it increased to 26.7. The results of the one-way ANOVA and LSD multiple comparison revealed that the difference in the average scores between Level 4 and Level 5 was not significant, whereas the differences between all the other adjacent levels were statistically significant (F(7, 163)=945.38, p<.05, MSe=21.44, p<.05).

To find out if the subjects' vocabulary sizes had differing effects, the subjects were then divided into five groups (2000, 3000, 4000, 5000, 6000-word levels) according to results of the TDU VLT online. Table 1 shows the number of the subjects, the mean and the total of the two sets of scores at each frequency level for each vocabulary size group.

As far as the mean scores of each JACET level were concerned, the groups with larger overall vocabulary size outperformed the lower vocabulary size groups at eight different levels without exception. Apart from the 6000-word level, scores decreased from Level 1 to Level 4 in accordance with the word frequency, as seen in Figure 2. However, beyond Level 4, the mean scores declined irregularly. In fact, the ANOVA and LSD multiple comparison showed that regardless of the vocabulary size of the subjects, the mean scores of the JACET level 4 and 5, 6

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Descriptive statistics for VLT scores by JACET level}
\end{figure}
Table 1. Descriptive statistics for VLT scores by vocabulary size

<table>
<thead>
<tr>
<th>Vocabulary Size</th>
<th>N</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>12</td>
<td>47.1</td>
<td>30.8</td>
<td>19.8</td>
<td>12.7</td>
<td>14.7</td>
<td>11.7</td>
<td>14.2</td>
<td>13.4</td>
<td>164.3</td>
</tr>
<tr>
<td>3000</td>
<td>38</td>
<td>51.3</td>
<td>38.9</td>
<td>28.0</td>
<td>20.5</td>
<td>19.6</td>
<td>18.7</td>
<td>19.6</td>
<td>17.4</td>
<td>214.1</td>
</tr>
<tr>
<td>4000</td>
<td>63</td>
<td>54.9</td>
<td>45.6</td>
<td>36.0</td>
<td>28.2</td>
<td>27.8</td>
<td>25.0</td>
<td>27.4</td>
<td>24.0</td>
<td>268.7</td>
</tr>
<tr>
<td>5000</td>
<td>43</td>
<td>57.8</td>
<td>54.2</td>
<td>45.5</td>
<td>35.2</td>
<td>35.9</td>
<td>31.8</td>
<td>33.0</td>
<td>30.4</td>
<td>323.8</td>
</tr>
<tr>
<td>6000</td>
<td>8</td>
<td>58.3</td>
<td>58.5</td>
<td>54.0</td>
<td>46.4</td>
<td>45.9</td>
<td>40.6</td>
<td>40.6</td>
<td>38.3</td>
<td>382.5</td>
</tr>
</tbody>
</table>

Figure 2. Mean values by estimated vocabulary size

and 7, and 7 and 8 were not significantly different (F(28,1113)=8.71, p<.01, MSe=18.6814, p<.05), with the exception of the 4000 vocabulary size group.

The above results show that the difficulty of Level 4 and Level 5 were almost the same. Furthermore, it is arguable that the smaller vocabulary-size groups such as 2000 and possibly 3000 levels suffered a floor effect caused by the test items being too difficult. The observed similar pattern of the decline or the lack of it in the mean scores of the larger vocabulary-size groups, however, indicate that it might not be the case. In short, we can safely conclude that the difficulty of the vocabulary level did not change much beyond the 5000 word-frequency band. In other words, the difficulty of the words could not be estimated by the frequency if the words belonged to bands beyond Level 4 of the JACET 8000.

4. Study 2

4.1 Purposes

The second phase was intended to discover if there are any discrepancies between learners' performance on the vocabulary test at each level and difficulty values judged by teachers of English.

4.2 Subjects
Forty English teachers (37 Japanese L1 and 3 English L1) participated in this study. They included experienced teachers as well as graduate students with high school teaching licenses. Their work settings were also rich in variety, ranging from junior high school teachers to university professors. By having various types of English teachers with different backgrounds, we believed the results of this questionnaire would be well balanced.

4.3 Method

An online Vocabulary Difficulty Questionnaire was developed for the purpose of this study (See Appendix 2). It is a web-based instrument, which asks teachers to click on a check box adjacent to each target word if they judge the word is “difficult” and to leave the box unchecked for a “not difficult” word (hereafter D and ND, respectively). The target words are the same 480 words used in Study 1. They were randomly distributed into six groups of 80 words, irrespective of their frequency levels. The subjects were presented with all six groups of words in random order. The instruction given to the subjects was “If you introduce those 480 words to students in two sessions by order of difficulty, which words would you choose as difficult words.” There was no time limit to complete the questionnaire.

4.4 Results

The basic statistics of the questionnaire results are shown in Table 2. The total number of clicks was 4486. In average, 112.15 words (SD =73.10) were judged as D words out of 480 words. Among the subjects, however, the difference in numbers of D words was very large, ranging from 289 to nil.

Figure 3 shows the number of words which teachers clicked as D words according to the different JACET levels. Interestingly, if we draw a line between the values of Level 1 and Level 8 as a graph of expected values, most of the actual scores for each level are plotted on this line.

To identify levels that manifested different behavior from the neighboring levels, a one-way ANOVA with LSD comparison was conducted. The results showed that only the differences between Level 3 and 4 and between Level 5 and 6 were statistically significant. (F(7,472)=44.29, p<.01) That is to say, the analysis showed that English teachers evaluated the words of Level 3 and Level 5 as being relatively easy compared to the words in neighboring bands.

To discover how many words were perceived as difficult in each frequency level, the number of words that were judged as difficult by more than three subjects was counted. Although we originally defined difficulty as “declared so by one or more subjects”, we could not eliminate the unintended mouse clicks for reasons mentioned below. Therefore, we excluded words that were

| Table 2. Basic statistics of the results of questionnaire for English teachers |
|-----------------------------|---------|--------|-------|-----|-----|
| N                           | Total Clicks | Mean ( /480) | SD    | Max | Min |
| 40                          | 4486       | 112.15     | 73.10 | 289 | 0   |
judged as difficult by less than 5% (two to be exact) of the subjects from this analysis. Figure 4 shows the results. At Level 1 and Level 2, the ND words outnumbered the D words, but beyond Level 3 there were more D words than ND words. At level 5 the number of D words reached 50 out of 60. When we look at the declining curve of the ND words, it is very similar to the declining curve of students' vocabulary test scores against frequency in Figure 1. As far as this analysis was concerned, we can conclude that the teachers' judgments of word difficulty were similar to the number of words correctly answered by students on the vocabulary test.

For the last stage of analysis, words judged as “difficult” by more than half of the teachers (20 teachers) were defined as D words. Figure 5 shows the number of D words by JACET level. Up to Level 3 there were no words chosen, but from Level 4 the number of D words increased gradually toward Level 8, apart from the sudden decline at Level 7. However, apart from Level 7, the difficulty of the words at each frequency band increased as the word frequency decreased. The difference between these and the previous results (shown in Figure 4) are striking. We can conclude that the teachers’ intuition of the difficulty of the words varied depending upon the method of analysis.
5. Discussion

These studies attempted to clarify the relationship between the frequency of a word and its difficulty. The results are tentative rather than conclusive, but there are four points to be discussed.

First of all, learners' overall mean scores on the TDU VLT declined at each frequency-band from Level 1 to Level 4, and beyond Level 4 they leveled off towards Level 8. This general conclusion supports Aizawa (2006). One possible explanation for this is that the first 4000 or 5000 words occupy a high lexical coverage of the texts used by students. A number of researchers have pointed out that L2 students can read texts reasonably well after acquiring 3000 word families (e.g. Nation, 2001; Nation & Waring, 1997). Laufer (1997) suggests that 3000 word families correspond to 5000 lemmas. If this is true, we should teach the first 5000 words to Japanese students as a core vocabulary.

The next point to be noted is that a group of students with higher English proficiency participated in the first experiment, including students with more than 6000 words. However, their mean scores for each frequency band followed the same pattern as other subjects; their score gradually declined up until Level 4, but did not change significantly between Level 4 and Level 5, or between Level 6 and Level 7. This result clearly shows that the score pattern of the subjects with higher proficiency followed the same pattern of the other subjects with lower proficiency. This suggests the possibility of estimating word difficulty by its frequency level up to Level 4.

Even so, the observed pattern needs to be interpreted with caution. JACET 8000 is regarded as one of the best vocabulary lists for Japanese EFL learners, and it is not our intention to criticize its validity. However, by observing the subjects' performance on the TDU VLT online, it became evident that none of the 60 test items from Level 5 retained an 80% correct rate while some items from other levels did. In fact, Level 5 test items included words such as bladder; regiment, thee, gastric, solicitor; and enzyme. Many of these words are not part of the subjects' daily vocabulary even in their native language. Nonetheless, the observed difficulty of the level by the teachers was not significantly different from that of Level 4. The conclusion to be drawn here is that Level 5 vocabulary items might actually be easier to learn than the ones in Level 4. If this is the case, the frequency level of vocabulary has little effect on difficulty. Further studies are needed to confirm this tentative assumption.

The last, but most important point to be discussed is that we obtained two conflicting results from the English teachers' questionnaire. If we take the total number of D words identified at each level, it was in accord with the frequency level. In other words, according to teachers' intuition, the more frequently a word is used, the less difficult it is for students. From this analysis, we can use the frequency of the word as an implicational scale for the difficulty of the word for students. On the other hand, if we take 50% teachers' agreement as a baseline, D words started to increase in number from Level 4 toward Level 8, except at Level 7. An
explanation for this is that the standard of 50% teachers’ agreement is not sensitive to the responses from a minority of teachers. In other words, teachers’ intuition of difficulty level varies a lot from teacher to teacher. To clarify this point, we need a further case study about the cutting off point for judgments.

6. Limitations and Implications

There are three implications for future studies. One is that we had 22 students with higher English proficiency from an English department, and this number of subjects may not seem to be enough. For future studies, it is desirable to include teachers or English specialists as subjects to find if this result can be extrapolated to other levels of EFL learners. Secondly, this study adopted two computer-based surveys, and in both cases subjects were asked to click the mouse. In this sort of measurement, certain factors have to be considered: unintended clicks of mice and fatigue from watching the screen and repeated simple actions. For a future study, a more sophisticated subject-friendly system needs to be developed. Finally, this study compared the difficulty of words by frequency band, not the difficulty of the individual word. Due to participants’ time constraints, we were reluctant to include too many target words. As a result, only 60 words were sampled randomly out of 1000 words for each JACET level. If we can measure the difficulty of words on a much larger scale, it will offer even deeper insights into the relationship between word frequency and its difficulty.

References


JACET English Lexicography SIG Workshop, Wayo Women's University.

Appendix

Appendix 1. Sample questions of TDU VLT online (Test 1, Level 1)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>center</td>
<td></td>
<td>2</td>
<td>goal</td>
<td></td>
<td>3</td>
<td>hole</td>
<td></td>
<td>4</td>
<td>police</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>success</td>
<td></td>
<td>4</td>
<td>example</td>
<td></td>
<td>5</td>
<td>rain</td>
<td></td>
<td>5</td>
<td>sense</td>
<td></td>
</tr>
</tbody>
</table>

Appendix 2. Sample items of Questionnaire (Form B)

INSTRUCTIONS

1. Click those words that you feel “DIFFICULT.”

2. When you are finished, please press “REVIEW” button.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>movie</td>
<td></td>
<td>2</td>
<td>bladder</td>
<td>☑</td>
<td>3</td>
<td>season</td>
<td></td>
<td>4</td>
<td>translation</td>
<td>☑</td>
</tr>
<tr>
<td>5</td>
<td>physiology</td>
<td>☑</td>
<td>6</td>
<td>recollection</td>
<td>☑</td>
<td>7</td>
<td>agenda</td>
<td></td>
<td>8</td>
<td>obscene</td>
<td>☑</td>
</tr>
<tr>
<td>9</td>
<td>terrace</td>
<td></td>
<td>10</td>
<td>sacrifice</td>
<td>☑</td>
<td>11</td>
<td>deceased</td>
<td></td>
<td>12</td>
<td>lifetime</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>cheerful</td>
<td></td>
<td>14</td>
<td>reproduction</td>
<td>☑</td>
<td>15</td>
<td>embarrassed</td>
<td></td>
<td>16</td>
<td>avoidance</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>perish</td>
<td></td>
<td>18</td>
<td>squire</td>
<td>☑</td>
<td>19</td>
<td>posture</td>
<td></td>
<td>20</td>
<td>expulsion</td>
<td>☑</td>
</tr>
</tbody>
</table>