The Effects of Teaching on Vocabulary Knowledge: Receptive vs. Productive

Kazumi AIZAWA, Natsue OCHIAI and Satsuki OSAKI  
Tokyo Denki University

Abstract

The gap between receptive and productive vocabulary seems to increase with progress in learning. It is estimated that Japanese EFL learners’ receptive vocabulary is double the size of their productive vocabulary (Tono et al., 1997). However, what is not so satisfactory from a pedagogical standpoint is empirical data to explain how this discrepancy happens. This study is designed to compare the effects of the two teaching methods that focused on either receptive or productive aspects of vocabulary at the two different proficiency levels. The target words were presented to students followed by basic receptive drills. Later these drills were supplemented either by receptive or productive drills. The result showed that learners in the production-centered class outperformed their counterparts in the reception-centered class in receptive tests a week later. However, in productive tests, the upper receptive and the upper productive groups did not differ in scores of the productive test. A possible explanation will be discussed for the effects of vocabulary teaching in the Japanese EFL classroom.

1. Introduction

Vocabulary knowledge has traditionally been classified into two aspects: receptive and productive. The terms “receptive” and “passive,” or “productive” and “active” are sometimes used interchangeably. However, because receptive skills include some active features (Nation, 2001), we use the terms “receptive” and “productive” (hereafter R and P) in this paper.

What is the relationship between L2 learners’ R and P vocabulary knowledge? Little is known about how large differences emerge between R and P vocabulary knowledge at each learning stage. At an elementary level, learners are introduced to a small number of new words. Teachers usually give them drills to practice using new words not only receptively but also
productively. However, more advanced learners usually have less opportunities for P practice. It seems likely that decreased opportunities to practice the P aspect of vocabulary at the later stage would enlarge the discrepancy between R and P vocabulary.

This study investigated the effects of R and P teaching on R and P aspects of vocabulary knowledge. We were interested in which teaching method was more effective for vocabulary learning for Japanese EFL university students.

2. Overview of the previous studies

2.1 Receptive vs. productive vocabulary knowledge

It is important here to define clearly the meaning of “receptive” and “productive” knowledge. Broadly speaking, there are three definitions. Nation (1990) defines “passive vocabulary knowledge” as the ability to recognize a word and recall its meaning when it is encountered, while “active vocabulary knowledge” is the ability to write the needed vocabulary as required. Meara (1990) suggests that active and passive vocabulary are “qualitatively different” in that passive vocabulary “can only be accessed if appropriate external stimulation is available” whereas active vocabulary “does not require any external stimulus, but can be activated by other words.” Melka (1997) proposes four stages of vocabulary; (1) imitation and/or reproduction without assimilation, (2) comprehension, (3) reproduction with assimilation, (4) production, but argues for a “continuum model” seeing boundaries between R and P as a degrees of knowledge or degrees of familiarity. In this paper, we will tentatively adopt the third view for reasons discussed later.

There are a limited number of studies that investigate teaching effects on learners’ vocabulary knowledge. Stoddard (1929) used a bi-directional translation task (English and French) and clarified three points; (1) R tests are easier than P tests, (2) the type of test favors the type of learning, and (3) the effect of the type of the test is greater than the effect of the type of learning. In other words, R tests are easier for both groups of learners than P tests, whichever teaching method (R or P) was used. Unfortunately, these conclusions were not based upon statistical analysis.

Mondoria (2002) compared two teaching methods that focused on either of R aspect or P aspect of vocabulary knowledge with Dutch learners of French. He found that a production-focused method yielded higher scores in both P and R tests.

In this study we further investigated the effects of direct teaching on vocabulary knowledge.

2.2 Measuring vocabulary knowledge

Several productive vocabulary tests for ESL learners have been constructed and used for a pedagogical and research purposes (See Table 1). We chose the Controlled Productive Levels Test (hereafter CPLT) for P knowledge and a translation task for R knowledge. There are two reasons for this. First, only the CPLT can quickly measure a certain set of words productively.
The rest of the P tests are designed for evaluating the written text and the word associations, which do not force learners to use certain words. Secondly, using CPLT together with a translation task makes it easier to measure two aspects of knowledge of the target words and allowed us to construct a set of parallel tests to measure R and P vocabulary knowledge.

Table 1. List of ESL productive vocabulary tests

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Test</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Productive</td>
<td>Lexical Frequency Profile</td>
<td>Laufer, 1994</td>
</tr>
<tr>
<td></td>
<td>Beyond 2000</td>
<td>Laufer, 1995</td>
</tr>
<tr>
<td></td>
<td>Controlled Productive Levels Test</td>
<td>Laufer &amp; Nation, 1999</td>
</tr>
<tr>
<td></td>
<td>Lex 30</td>
<td>Meara &amp; Fitzpatrick, 2000</td>
</tr>
</tbody>
</table>

3. Study

3.1 Purposes

The purpose of this study was to compare the effects of R and P teaching on vocabulary acquisition. To do this, we taught two groups of Japanese EFL learners; one using a reception-centered and the other using production-centered method, and measured their vocabulary knowledge by the two methods to find which group would score higher. The hypotheses were:

1. Learners score higher when tested in the same method as taught
2. When learners are taught productively, their total scores of the R test and the P test will be higher than those scores when taught receptively
3. There are differences in scores of the R test and the P test depending upon learners’ proficiency levels or the teaching method

3.2 Method

3.2.1 Subjects

The subjects were 80 Japanese university students from four different classes in an engineering department. They were taking an English class as part of their requirements. Thirty-six of them (18 × two classes) were in upper level classes and 44 (22 × two classes) were in lower level classes. One class at each level was taught using reception-centered method, and the other production-centered one. The four groups were distinguished as “upper R,” “upper P,” “lower R” and “lower P” ones.

3.2.2 Material

Based upon a pilot study with learners of equivalent proficiency levels, we selected 20 items from CPLT (version 1 and 2), 17 items from Academic Word List, and three items from 3000-word level. The two criteria for this selection were (1) fewer students knew the target words, and (2) the target words are not used as loan words in Japanese. The items were divided
into two groups: List 1 and List 2 (See Table 2). In addition to the target words, two distracters were added to minimize the priming effect. Therefore, each list comprised of twelve items. To ensure that test takers understand the sentences in questions, Japanese glosses for difficult words in sentences were provided. For the R test, new sentences using the target words were given as a translation task. The P test had an identical format except that learners answered in English (See Appendix A and B).

Table 2. Target words

<table>
<thead>
<tr>
<th>List 1</th>
<th>AWL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>intimacy, motive, subsequent, anomaly, intellect, indicate, assess, restore, rational</td>
</tr>
<tr>
<td>3000-word level</td>
<td>lawn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>List 2</th>
<th>AWL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>doctrine, inspect, accumulate, evaluate, attain, inherent, sequence, ensure</td>
</tr>
<tr>
<td>3000-word level</td>
<td>vein, supreme</td>
</tr>
</tbody>
</table>

Both of the R and the P tests had two versions with ten questions and two dummies. Each item was assigned two points and the maximum possible score for each test was 20. To ensure that learners were not familiar with the words, a revised version of Self-Reported Knowledge Scaling (Barfield, 2001) was also prepared. There were 20 target words and ten dummy words ordered alphabetically. Each correct answer counted as two points, making the total 40. Sample questions are shown in the Appendix C.

3.2.3 Procedure of teaching

The teaching plans for R and P classes are summarized in Table 3. These procedures presupposed that learners could acquire the target words either at the “comprehension level” or at the “reproduction with assimilation level” by Melka’s definition. In R teaching, learners were introduced to new word meaning and pronunciation using flashcards, followed by the translation task. Mechanical exercises were supplemented by activities on a worksheet. Learners were asked to answer the word meaning in a new context. These two activities took about 15 minutes.

Table 3. Teaching procedure for a session

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Receptive Teaching</th>
<th>Productive Teaching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Imitation → Comprehension</td>
<td>Imitation → Reproduction with assimilation</td>
</tr>
<tr>
<td>Time Alotted</td>
<td>Activities</td>
<td>Material</td>
</tr>
<tr>
<td>7 minutes</td>
<td>1. Mechanical exercises</td>
<td>English into Japanese, English into Japanese, Flashcard</td>
</tr>
<tr>
<td></td>
<td>2. Exercises of worksheet</td>
<td>English into Japanese, Japanese into English</td>
</tr>
<tr>
<td>8 minutes</td>
<td></td>
<td>Worksheet</td>
</tr>
</tbody>
</table>
In P teaching, we used two different activities (shaded area in Table 3). First, learners were asked to spell a word by looking at the Japanese translation on a flashcard. Secondly, learners were asked to write the words in the blanks in sentences on the worksheet.

3.2.4. Treatment

In the first week, we handed learners Self-Reported Knowledge Scaling to check their knowledge of the target words. This was followed by 15 minutes of vocabulary teaching. On the following week, students were given a P test, and then an R test at the beginning of the lesson. We believed that if learners were given the P test first, there would be little interference in the R test. These tests were given as surprise tests. During the same session, 15 minutes before the lesson end, ten new words were introduced. The same teaching procedure was followed as before. On the third week, the words introduced on the second week were tested first productively and then receptively.

3.3 Result

For the first stage of analysis, we analyzed subjects' scores in TOEIC to verify the equality in English proficiency of the pair groups: two upper and two lower groups. The mean scores of the upper P group and the upper R group were 451.9 and 431.4 respectively. On the other hand, the lower P group marked 318.0 where the lower R group scored 297.3. A two-tailed t-test showed that no significant difference was found either between upper proficiency groups (t(34)=0.45, n.s.) nor the lower proficiency groups (t(42)=0.78, n.s.). Based upon these outcomes, the pair groups were considered to be equal in proficiency. Likewise, we analyzed the scores of the Self-Reported Knowledge Scaling to validate the equality of subjects' knowledge of the target words. The result showed that at the upper level, the mean score of the P group was 2.2 while the R group scored 1.4. At the lower level, the P group marked 0.3 whereas the receptive marked 0.4. The result of the two-tailed t-test showed that there was no significant difference either between the upper groups (t(34)=0.25, n.s.) nor the lower groups (t(42)=0.78, n.s.). From these results, we assumed that the subjects did not know the target words.

For the second stage, we compared the scores of the R and the P tests by two groups: R or P. In the R test, the mean score of the R group was 17.53 and that of the P group was 19.23 (See Table 4). However, there was no significant difference between the two groups (t(78)=1.17, n.s.). On the other hand, in the P test, the R group scored 16.00 and the P group scored 21.40. A two-tailed t-test showed that a significant difference was found between two groups (t(78)=3.39, p<.05). Therefore, the production-centered method was more effective on the P knowledge than the reception-centered method and so we partly supported the first hypothesis: learners score higher when tested in the same method as taught in the P group.

For the third stage, the total average scores of the R and the P tests were compared between the pair groups. As can be seen from Table 4, the mean score of the R group was 33.53 and that
Table 4. Basic statistics for the vocabulary test at two different groups

<table>
<thead>
<tr>
<th>Testing Method</th>
<th>Receptive (SD)</th>
<th>Productive (SD)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Receptive (N=40)</td>
<td>17.53 (8.87)</td>
<td>16.00 (9.93)</td>
</tr>
<tr>
<td></td>
<td>Productive (N=40)</td>
<td>19.23 (9.55)</td>
<td>21.40 (9.69)</td>
</tr>
</tbody>
</table>

(40 points possible)

of the P group was 40.63. A two-tailed t-test showed that there turned out to be marginally different between them (t(78)=1.96, p<.10). It would be possible to conclude that the production-centered method was more effective to learner’s vocabulary knowledge than the reception-centered method. Therefore, we supported the second hypothesis: When learners are taught productively, their total score of the R and the P tests will be higher than when taught receptively.

![Figure 1. Comparison of the means between each level class](image)

For the forth stage of analysis, the scores of four groups in each type of the tests were compared (See Figure 1). As for the R test, a one-way ANOVA revealed that there was significant difference between the four groups (F(3, 76)=7.59, p<.01). A further analysis of multiple comparisons by LSD revealed that the scores of the upper P class was significantly higher than those of the other groups (MSe=69.47, p<.05). However, there were no significant differences among the rest of the three groups. In case of the P test, a one-way ANOVA revealed that there was a significant difference among the four groups (F(3, 76)=7.68, p<.01). An analysis of LSD showed that the scores of the upper P, the upper R, and the lower P groups were significantly higher than the lower R class (MSe=83.67, p<.05), but there was no significant difference among the three groups. This result allowed us to confirm the third hypothesis: There
are differences in scores of the R and the P tests depending upon learner’s proficiency level or the teaching method.

3. Other findings

In order to investigate the pattern in errors and misuses in the P test by the four groups, we assumed that every item in the P test was answered by a different subject. This assumption allowed us to analyze the data by a Chi-square test (See Table 5). In this table, shaded area means “misuse (partly correct),” while white area stands for “error (incorrect).” The dark triangles show significantly higher cells while the white inverted triangles means significantly lower ones (p<.05). There were five significantly higher cells and three significantly lower cells. The characteristics of errors in each group can be summarized as follows:

1. In the lower P group, relatively many misuses in the inflection and the tense were observed (36 misuses).
2. In the lower R group, there was the largest number of blanks of the four groups (203 blanks).
3. In the upper P group, blanks were the fewest (54 blanks), while misuses in spellings or in parts of speech were relatively many (26 and 15 misuses).
4. In the upper R group, subjects tended to answer with the wrong words starting with the same letters given as hints (45 errors).

Table 5. Analyses of misuses and errors in the P test

<table>
<thead>
<tr>
<th></th>
<th>spelling</th>
<th>part of speech</th>
<th>inflection and tense</th>
<th>wrong word starting with same letters</th>
<th>spelling</th>
<th>blank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Productive</td>
<td>25</td>
<td>6</td>
<td>36 ▲</td>
<td>27</td>
<td>38</td>
<td>115</td>
</tr>
<tr>
<td>Lower Receptive</td>
<td>37</td>
<td>5 △</td>
<td>14 △</td>
<td>46</td>
<td>46</td>
<td>203 ▲</td>
</tr>
<tr>
<td>Upper Productive</td>
<td>26 ▲</td>
<td>15 ▲</td>
<td>19</td>
<td>25</td>
<td>14</td>
<td>54 △</td>
</tr>
<tr>
<td>Upper Receptive</td>
<td>25</td>
<td>8</td>
<td>15</td>
<td>45 ▲</td>
<td>21</td>
<td>119</td>
</tr>
</tbody>
</table>

4. Discussion

This study attempted to clarify the teaching effects of the reception-centered or the production-centered method on the R or the P knowledge. The results were tentative rather than conclusive, but there are three points to be discussed.

First of all, the production-centered teaching method was found to be more effective on the overall vocabulary knowledge than the reception-centered one. In case of the lower proficiency learners, the reception-centered treatment turned out to be significantly lower in the P test score. Therefore, it may be necessary to incorporate some P drills to develop their P knowledge. In the case of the upper proficiency learners, the production-centered treatment improved learners’ R
knowledge as well. A possible explanation for this is that if the learner did not have enough proficiency, they could not use the words productively without a P treatment. Therefore, we can recommend that it is desirable to incorporate some productive drills for the new words, especially for low proficiency learners. In our study, the mean score in TOEIC by the lower level learners was about 300. This might give us insightful suggestions that it may be particularly useful for the learners of this level of proficiency to have a direct treatment of the P vocabulary.

Secondly, when learners' proficiency was higher, the P treatment contributed to improve the R knowledge as well. This might be because the learners' level of processing improved by practicing the words productively. This must have resulted in more retention of the words at the R level. The P treatment seems to have a ripple effect on the R knowledge at the upper level.

However, the upper level learners did not differ in P test scores in whichever teaching method they were taught. There are two possible explanations for this. Firstly, learners might have enough proficiency to acquire the words at the "reproduction with assimilation level" with treatment as R vocabulary. If this interpretation is possible, we can deduce that there is a certain proficiency level at which learners can use the target word productively without any P treatment. This assumption would explain the classroom observation that there are certain cases in which learners can use new words productively without drills in speaking or writing. Actually, it is physically impossible for teachers to give learners drills for every single word productively. An alternative explanation is that the P test did not measure only the P aspect of vocabulary knowledge. To answer correctly in the CPLT, learners needed to have a receptive vocabulary and reading proficiency to understand the sentences used for each question.

Thirdly, the error analysis of the productive test brought us interesting findings. When the lower proficiency learners received the R treatment, the blanks were by far the largest in number. On the other hand, when the upper level learners received the R treatment, they tended to choose the wrong words starting with the same letters. We believe that those responses were caused by the reception-centered treatment. This result also seemed to validate the effects of production-centered teaching on the P aspect of vocabulary knowledge.

In the production-centered groups, minor errors outnumbered those in the reception-centered groups. This is probably because the learners in those groups did not have chance to review the target words after the treatment. If they had had time to review the new words after the session, they might have elaborated their retention of the words, improving the P test scores considerably.

5. Limitations and implications

The main limitations of this study stems from the ambiguous nature of the CPLT as a test. As Read (2000) points out, it is not clear what the CPLT actually measures. For example, there is a difference in the amount of knowledge required for learners to answer to each item since the number of letters used as hints in each blanks differs. Clearly further research is needed to

158
establish the reliability and validity of the test as a measure of L1 knowledge.

Another limitation is that the subjects in this study were Japanese university EFL students who had the similar levels of English proficiency. Further investigations are needed to determine whether the conclusions in this study can be extrapolated to different levels of learners. If we can get the same results as this study, it will be a step closer to a more generalized model of vocabulary knowledge.

The principle pedagogical implication is that there is a need to develop a systematic program for vocabulary teaching according to learners’ proficiency levels. Even when a word is taught using the same method either receptively or productively, the effect of the teaching will vary depending on the learners’ proficiency levels.

References


Appendix

A. Sample Questions of Receptive Test
例にならって、下線部の語の意味を日本語で答えなさい。

(例) I use cream and sugar in my coffee. [砂糖]
3. I enjoyed walking across the green lawn.
4. The experience created an increased intimacy between us.
5. She possesses the intellect and the energy needed for the job.
6. The effect on the environment should be carefully assessed before building starts.
7. There is no rational explanation for going out on such a rainy day.

B. Sample Questions of Productive Test
例にならって、与えられた単語で始めまる単語を入れて句を完成しなさい。

(例) The birds were singing in the trees. [singing]
3. The story tells us about a crime and subsan punishment.
4. Many people in England mow the lawn of their houses on Sunday morning.
5. The new manager’s job was to restore the company to its former profitability.
6. It’s difficult to assess a person’s true knowledge by one or two tests.
7. Spending many years together deepened their intimacy.

C. Samples of Self-Reported Knowledge Scaling
次の単語について、次のどれに当てはまるか答えてください。ただし、(3)と(4)の場合はその意味も答えてください。

(1) 見たこともないし、見当もつかない
(2) 見たことはあるが、意味はわからない
(3) 見たことはありそうだが、考えてみないと意味はわからない
(4) 考えなくても意味がわかる

<table>
<thead>
<tr>
<th>番号</th>
<th>意味</th>
<th>番号</th>
<th>意味</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>accumulate</td>
<td>16</td>
<td>ensure</td>
</tr>
<tr>
<td>2</td>
<td>anomaly</td>
<td>17</td>
<td>evaluate</td>
</tr>
</tbody>
</table>