Second Language Incidental Vocabulary Retention: The Effect of Text and Picture Annotation Types

Makoto Yoshii
Baiko Gakuin University

Jeffra Flaitz
University of South Florida

ABSTRACT
The present study examines the effect that annotation type has on L2 incidental vocabulary retention in a multimedia reading setting. Three annotation types were compared: text-only, picture-only, and a combination of the two. The participants were 151 adult ESL learners at beginning and intermediate language proficiency levels. The participants read a story for comprehension purposes using the Internet. Three types of instruments were used for vocabulary retention assessment: Picture recognition, Word recognition, and Definition Supply tests. The results ANOVA analyses indicate that the Combination group (annotations with text and picture) outperformed the Text-only and Picture-only groups on the immediate tests. The Combination group also outperformed the other two groups on the delayed tests, however, the differences were smaller than those for the immediate tests. There was no significant interaction between annotation type and proficiency level for either the immediate or the delayed tests. Repeated measure ANOVAs revealed no significant differences among the groups in the rate of change between the immediate test scores and delayed test scores. The participants’ scores on the delayed tests, regardless of the group to which they were assigned, declined equally from those of the immediate tests.

KEYWORDS
Annotation Studies, Incidental Vocabulary Learning, ESL, Experimental Study, Web-based Reading Activity

INTRODUCTION
This study examines the effectiveness of annotations on incidental vocabulary learning (retention) in an L2 reading program for adult ESL students. The term Incidental is defined here as learners’ acquisition of the meanings of words as they engage in other tasks such as comprehension of reading and listening passages. Huckin and Coady (1999) defined incidental learning, particularly that dealing with reading, as “a by-product, not the target, of the main cognitive activity, read-
The Effect of Text and Picture Annotation on Vocabulary Retention

ing.” The study here extends the current research on the topic (Hulstijn, 1992; Jacobs, Dufon, & Fong, 1994; Scherfer, 1993; Watanabe, 1992) by placing the task in a multimedia setting. Previous research has dealt little with vocabulary learning in a technology-enhanced environment. Moreover, the research to date has been limited to a single L1 group studying a second language, but no study has looked at the effect of annotations on L2 vocabulary learning in a multilingual group setting (different L1s). In addition, participants in previous studies have generally been either intermediate or advanced students. Few researchers have examined the incidental vocabulary retention of novice L2 learners.

Research design and materials have also proven problematic in studies of this type. For example, studies by Chun and Plass (1996) and Plass, Chun, Mayer, and Leutner (1998) demonstrated the superiority of a combination of Text-and-Picture over other types of glossing for incidental L2 vocabulary learning in multimedia settings, but these studies were conducted with a within-subjects design, and the results require confirmation by between-subjects design studies. Chun and Plass (1996) pointed out that their study compared Text-only versus Text-and-Picture or Text-and-Video and that there is a need for studies to investigate single gloss types. Kost, Foss, and Lenzini (1999) conducted a study focusing on a between-subjects design and further differentiated the variables (Text-only, Picture-only, and Text-and-Picture). That study, however, made use of a printed text and was, therefore, quite different in format from that using a multimedia setting. For example, in a printed text, glosses appear in the margin of the text; in this way, students may be exposed to the glosses at all times. In a multimedia setting, on the other hand, glosses are available to learners only upon their request, and students generally access only one gloss at a time.

The current study attempts to address these design shortcomings. It also takes the learner’s language proficiency into consideration with regard to the effectiveness of annotation type. Two proficiency levels—beginning and intermediate—are included in the study.

BACKGROUND

Vocabulary learning is an essential part of language learning. Learning words can considered to be the most important aspect of second language acquisition (Knight, 1994). Candlin (1988) stated that “... the study of vocabulary is at the heart of language teaching in terms of organization of syllabuses, the evaluation of learner performance, and the provision of learning resources....” Maiguashca (1993) said that vocabulary is “perhaps the fastest growing area of second language education in terms of research output and publication.”

Using the framework given by Coady (1997) and Hulstijn, Hollander, and Greidanus (1996), Hunt and Beglar (1998) identified three approaches to enhance vocabulary learning, namely, incidental learning, explicit instruction, and independent strategy development. Among the three, incidental vocabulary learning was viewed as an essential part of L2 vocabulary acquisition. Nation (1999) stressed
the importance of incidental learning through “message-focused activities” as follows: “A well-balanced language learning programme has an appropriate balance of opportunities to learn from message-focused activities and from direct study of language items, with direct study of language items occupying no more than 25% of the total learning programme.” According to Huckin and Coady (1999), many studies seem to indicate “except for the first few thousand most common words, vocabulary learning dominantly occurs through extensive reading, with the learner guessing at the meaning of unknown words.”

There are, however, disadvantages for incidental vocabulary learning compared to more direct learning. Incidental vocabulary learning is not always effective or efficient (Hulstijn, 1992; Mondria & Wit-de Boer, 1991; Nation, 1982). Researchers point out that contextual information is often ambiguous and not sufficiently reliable for L1 and L2 learners to be able to make the correct inference (Bensoussan & Laufer, 1984; Hulstijn, 1992; Mondria & Wit-de Boer, 1991). Learners run the risk, therefore, of failing to verify the correctness of inferences and can learn words incorrectly (Carnine, Kameenui, & Coyle, 1984; Dubin & Olshtain, 1993; Huckin & Haynes, 1993; Hulstijn, 1992; Mondria & Wit-de Boer, 1991).

How, then, can these disadvantages to incidental vocabulary retention be overcome? Hulstijn, Hollander, & Greidanus (1996) identified the use of marginal glosses as one way to enhance incidental vocabulary learning. Studies have shown that incidental L2 vocabulary learning is enhanced by the use of glosses in printed materials (Hulstijn, Hollander, & Greidanus, 1996; Jacobs, Dufon & Fong, 1994; Watanabe, 1997). For example, Jacobs, Dufon, & Fong (1994) worked with American students studying Spanish as an L2 in a fourth semester course. They found that learners who had access to glosses outperformed those who did not have glosses on a vocabulary test administered immediately after treatment. Hulstijn, et al. (1996) conducted research with Dutch students learning French as an L2 and found that marginal glosses (L1) were more effective than bilingual dictionary use or a Text-only condition (no glosses and no use of dictionary). Watanabe (1997) investigated how text modification and task would affect incidental vocabulary learning. Japanese university students retained less English vocabulary when they worked with texts containing no modifications, but more English vocabulary when they worked with texts containing L2 glosses. L2 glosses were also more effective than appositives—simplified restatements of difficult words inserted immediately after the words in question in the text.

Studies have also examined the effects of different types of glosses on incidental L2 vocabulary learning (Jacobs, Dufon & Fong, 1994; Kost, et al., 1999; Watanabe, 1997). Jacobs, et al. (1994) compared the effectiveness of L1 glosses and L2 glosses but did not find any significant difference between the effectiveness of the two. Comparing regular and multiple-choice glosses in which students were asked to choose the correct definition from two alternatives offered, Watanabe (1997) found no significant difference in effectiveness between either of the two annotation types. Kost, et al. (1999) compared three types of glosses: Text-only...
The Effect of Text and Picture Annotation on Vocabulary Retention

(L1), Picture-only, and Text-and-Picture. The study was conducted with American students who were studying German as an L2 in a second semester course. The combination of Text-and-Picture glosses was the most effective of the three types.

In addition to the above-mentioned studies, there are a few cases that have examined the effect of glossing with computerized materials. Chun and Plass (1996) and Plass, et al. (1998) conducted a series of experiments examining the effect of different types of multimedia glosses for American university students who were studying German as an L2 in intermediate level classes. These studies had a within-subjects design, and students had the option of choosing from the different types of glosses: text (L1), pictures, and video. The results showed that students were able to remember more vocabulary words when they had accessed the combination of text and pictures.

Nagata (1999) conducted an experiment which was basically a computerized version of Watanabe’s (1997) study. She compared the effectiveness of a single gloss and a multiple-choice gloss as American college students taking the second semester Japanese course read a text on the computer. She found that the multiple-choice gloss was significantly more effective than the single gloss.

RESEARCH QUESTIONS

Just as Nagata (1999) conducted the computerized version of the experiment following Watanabe’s (1997) printed text experiment, there is a need to conduct a study similar to that of Kost, et al. (1999) to examine the effect of different annotation types while adapting the study to a multimedia setting.

The main research question for this study is: What effect do multimedia annotations have on the incidental vocabulary learning of ESL students when they read a story for comprehension? The investigation here addresses the following specific questions:

1. Does a particular annotation type affect immediate vocabulary retention when learners read a text for comprehension purposes?
2. Does a particular annotation type affect delayed vocabulary retention when learners read a text for comprehension purposes?
3. What is the degree of interaction between L2 proficiency level and annotation type on tasks involving immediate recall of vocabulary?
4. What is the degree of interaction between L2 proficiency level and annotation type on tasks involving delayed recall of vocabulary?
5. Is there a statistically significant change in learners’ vocabulary recall over time (between the immediate and delayed tests) according to the types of annotations used?

METHOD

Participants

Participants (n = 151) were ESL students enrolled in English language institutes...
at universities in Florida. Beginning and intermediate level students, as defined by the individual universities, participated in the study. Five intensive English programs were selected because they had well established programs and shared similar course objectives and curricula. These universities provide ESL classes at a variety of proficiency levels, ranging from beginning to advanced. The participants represented 38 countries and 18 languages. There were 69 female students and 82 male students. Their average age was 24.6, ranging from 16 to 47 years old.

**Dependent and Independent Variables**

The primary independent variable was the type of multimedia annotation in the reading material on the web. There were three types: (a) Text-only, (b) Picture-only, and (c) Text-and-Picture. Figure 1 shows an example of the annotation types. A graphic artist, upon the request of the researcher, created the pictures.

![Figure 1](image)

**Instruments**

**Pretest**

The pretest, which included 14 target words and 10 additional distracters, served to verify the participants’ lack of familiarity with the target words. Participants were instructed to put a check mark by a word they knew and to provide a brief written explanation in either the L2 or the L1. A cutoff point was set at 30% of the total number of target words to identify and remove from the study any exceptional students with high pretest scores.

**Reading Material and Target Words**

A short story, “A Scary Night,” created for ESL students by the researcher, was used for the study. (See the reading material in Appendix A.) There were 14 target words in this study, each of which had to meet several conditions to be selected. First, they had to be noncognate words for the learners. As a means of
verification, Spanish, Japanese, French, German, and Arabic speakers checked and confirmed that the words were not cognates (e.g., “preoccupy” in English resembling *preocupar* in Spanish). Many previous annotation studies had dealt mainly with nouns, and some studies had dealt with multiple parts of speech. In this study, the target vocabulary words were verbs.

The choice of words partially depended upon whether or not they could be effectively drawn as pictures. After pictures were drawn to represent each target word, they were validated both by several instructors and students, who were not participants in the study, in order to make certain that the pictures conveyed the intended meaning.

Posttests

Two posttests—an immediate and a delayed (2 weeks) series of measures were administered to the participants. Each posttest consisted of three parts common to both tests: Definition Supply (in either the L1 or the L2), Picture Recognition, and Word Recognition. The formats of these tests were similar to the format used in Kost, et al.’s (1999) study.

Definition Supply Test

All 14 target words appeared in the test. Students were asked to put a check mark by a word they remembered and to supply the meaning of the word in either the L1 or the L2. (See Definition Supply test format in Appendix B.)

One point was given to a partially correct answer (lenient evaluation), and two points were given to a fully correct answer (strict evaluation) according to the scoring criteria set for this study. (See the detailed evaluation criteria in Appendix C.) Otherwise, the students did not receive any points.

Two raters were assigned to each language group, and interrater reliability was established for the immediate and the delayed Definition Supply tests. The raters were either native speakers of the language or had near-native proficiency.

Picture Recognition Test

In the Picture Recognition tests, students had to choose from four pictures to identify the drawing which best conveyed the meaning of a given verb. (See Picture Recognition test format in Appendix D.) One point was given to a correct answer. All of the pictures in the test were different from the ones used in the glosses, but their content conveyed the same meaning. This safeguard was taken to avoid random guessing by students who may have remembered an annotation picture they saw in the reading without understanding its meaning. These pictures were taken from the “Royalty-Free Clip Art Collection for Foreign/Second Language Instruction” at Purdue University (www.sla.purdue.edu/fll/JapanProj/FLClipart/default.html). This collection was chosen since the pictures were “simple line drawings designed to be as culturally and linguistically neutral as possible for foreign language instruction” (Kost, et al., 1999).
Word Recognition Test

All 14 target words were used for the Word Recognition test for the same reasons as explained above for the Picture Recognition test. (See Word Recognition test format in Appendix E.) In this test, students had to select the definition of a given verb from four choices. One point was given to a correct answer. Among the four choices, one was the correct answer, another was incorrect, and the other two were distracters taken from the context surrounding the word in question in the reading material. All of the definitions (i.e., correct answers) came from the story but were phrased differently from those used in the glosses. All the definitions were written in the L2 (English).

Treatment

The experiment was conducted during the students’ regular class times, and required two consecutive 50-minute sessions. The participants were stratified by L1s in each class at each level and were randomly assigned beforehand to one of the conditions: Text-only, Picture-only, and Text-and-Picture. A number was assigned to each student, and cards containing their assigned number were distributed to the students.

In the first class, the participants filled out a consent form and were asked to take a vocabulary pretest. Next, the students filled out a questionnaire. This order was implemented in order to divert students’ attention from the vocabulary item before the treatment was begun. The questionnaire asked for students’ demographic information such as gender, age, and extent of English language study before and after arriving in the US. The questionnaire also asked for students’ perceptions and attitudes toward computers as well as their familiarity with their use.

In the second class hour, students went to a lab where they were asked to go to numbered computers that matched the assigned numbers on their cards, one student per computer. A brief oral introduction was given to the web-based reading activity. After the students went through the tutorial, they were told to click on the small computer icon at the bottom of the page. After entering their identification numbers, the students were asked to read the story. They were also informed that they would answer comprehension questions and complete a short survey concerning their impressions of the computerized reading material. They were not told that they would take a posttest involving vocabulary. Participants read the story, and, when they finished the reading, they were directed to close the web site. They were then asked to raise their hands in order to receive the comprehension questions. As individuals completed the reading, they were given a brief comprehension task. When the comprehension task was complete, the students were again asked to raise their hands. This first postreading task was collected, and, at this point, the participants were asked to complete the Definition Supply test. The same procedure of the students individually raising their hands, submitting their tests, and receiving another unexpected task was again
performed for the Picture Recognition test followed by the Word Recognition test. Finally, students were given the posttreatment survey of which they had been previously informed. Two weeks later, students were again unexpectedly given delayed posttests. The content of these tests was the same as for the immediate posttest, minus the comprehension questions and the survey, though the order of the items in each section differed from that of the previous tests.

**Data Analysis**

Each research question was analyzed by means of ANOVA. A 3 x 2 ANOVA (type of annotation x language proficiency level) was performed on students’ scores on the immediate tests and the delayed tests. The data were examined in terms of the effect of the type of annotation (Text-only, Picture-only, and Text-and-Picture), the effect of language proficiency level (beginning or intermediate), and the interaction effect of the two (the type of annotation and language proficiency). Post hoc analyses were applied to further examine the differences between the groups.

A mixed model ANOVA was also conducted in order to get an overall picture of the effects of annotation types and the changes that occurred from the immediate to the delayed posttests. This analysis involved one between-subjects (the type of annotations) and one within-subjects (immediate and delayed) test. The analyses involved the investigation of group differences on the immediate and delayed tests. They also investigated whether the differences among the groups changed over time between the first and the second posttests.

The alpha level was set at .05 for all the analyses. When significant differences emerged, the contrasts between the pairs of least square means of the groups were evaluated. In order to control for error, a Holm modified Bonferroni procedure was used: The alpha level for the first comparison between the two groups was set at .0167 (one third of .05), that of the second comparison at .025 (half of .05), and that of the third comparison at .05.

**RESULTS**

**Pretest Results**

The scores on the pretest (Definition Supply format) were first analyzed in order to check the equivalency of the three condition groups in terms of the students’ pre-knowledge of the target words. The cutoff point was set at fewer than five correct words (30%) out of 14 target words. Only two students were removed from the study because their scores exceeded the cutoff point (one with a score of 5, and the other 6).

A two-way ANOVA (Group x Level) was performed for scores derived under both the strict and the lenient assessment conditions. The significance level was set at .05. The results indicated that there were no significant differences among the three groups, either for the strict scoring, $F(2, 145) = 0.49, p > .05$ or for the lenient scoring, $F(2, 145) = 0.12, p > .05$. The ANOVA results indicated
that the participants in the three condition groups were equivalent in terms of their performances on the Definition Supply pretest. The average mean score on the pretest was .29 under the strict scoring conditions and .51 under the lenient scoring conditions, indicating that the average student knew less than one word out of 14 target words prior to the study.

**Posttest Results by Research Question**

**Question 1**

Effect of Annotation Type for Immediate Vocabulary Retention

The first research question inquired whether a particular annotation type might affect immediate vocabulary retention when learners read a text for comprehension purposes. In order to answer this question, the results of the analysis of the three immediate tests (Picture Recognition, Word Recognition, and Definition Supply—under both the strict and lenient scoring conditions) were examined (see Table 1).

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M/SD</th>
<th>M/SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>50</td>
<td>7.46</td>
<td>2.53</td>
<td>7.58</td>
<td>2.60</td>
<td>1.86/1.80</td>
<td>3.52/2.24</td>
</tr>
<tr>
<td>Picture-only</td>
<td>51</td>
<td>7.18</td>
<td>2.98</td>
<td>6.41</td>
<td>2.87</td>
<td>0.96/1.39</td>
<td>3.00/2.39</td>
</tr>
<tr>
<td>Text-only</td>
<td>50</td>
<td>5.98</td>
<td>2.48</td>
<td>6.12</td>
<td>3.05</td>
<td>1.38/1.63</td>
<td>2.62/2.46</td>
</tr>
</tbody>
</table>

As Table 1 shows, the Combination group outperformed the other two groups on all three test types. Significant differences were found among the groups for the Picture Recognition test, $F(2, 145) = 4.04, p < .05$. The contrasts revealed that the Combination group outperformed the Text-only group significantly ($p < .0167$). There were also significant differences among the groups for the Word Recognition test, $F(2, 145) = 3.28, p < .05$. The contrasts, however, did not reveal any significant pairwise differences among the groups. For the Definition Supply test, there were significant differences under the strict scoring conditions, $F(2, 145) = 3.93, p < .05$. The contrasts revealed a significant difference solely between the Combination and the Picture-only groups ($p < .0167$). There were no significant differences among the groups for the Definition Supply test under the lenient scoring conditions. Table 1 also shows that the Picture-only group held second place on all tests, except for the Definition Supply test evaluated under the strict scoring conditions. However, the contrasts did not reveal any significant difference between the Picture-only and the Text-only groups for any of the immediate tests.
Question 2
Effect of Annotation Type for Delayed Vocabulary Retention

The second research question examined whether a particular annotation type might affect delayed vocabulary retention when learners read a text for comprehension purposes. In order to answer this question, students’ scores on the three delayed tests (Picture Recognition, Word Recognition, and Definition Supply—under both the strict and lenient scoring conditions) were examined (see Table 2).

Table 2
Means and Standard Deviations of the Delayed Test Scores by Group

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Picture Recognition</th>
<th>Word Recognition</th>
<th>Definition Supply</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Combination</td>
<td>50</td>
<td>6.48</td>
<td>2.67</td>
<td>6.06</td>
</tr>
<tr>
<td>Picture-only</td>
<td>51</td>
<td>6.31</td>
<td>3.08</td>
<td>5.43</td>
</tr>
<tr>
<td>Text-only</td>
<td>50</td>
<td>4.92</td>
<td>2.78</td>
<td>4.62</td>
</tr>
</tbody>
</table>

Table 2 indicates that the Combination group outperformed the other two groups on all measures, except for the Definition Supply test under the lenient scoring condition. The group differences were significant for the Picture Recognition test, $F (2, 145) = 4.17, p < .05$. The contrasts revealed a significant difference between the Combination and the Text-only groups ($p < .0167$) and between the Picture-only and Text-only groups ($p < .025$). There were no significant differences among the groups for the Word Recognition test or the Definition Supply test (strict and lenient).

These results seem to indicate that the Combination group again was most effective in terms of delayed vocabulary retention. The differences among the groups, however, were smaller on the delayed test than on the immediate test.

Question 3
Interaction between Level and Group for the Immediate Recall Tests

The third question examined the degree of interaction between proficiency level and annotation type on tasks involving immediate recall of vocabulary. In order to answer this question, all the interaction effects between Level and Group for the three instruments were examined. Table 3 lists the interaction effect between the two main effects for the three immediate tests.
Makoto Yoshii and Jeffra Flaitz

Table 3
Summary of Interaction between Group and Level for the Immediate Tests

<table>
<thead>
<tr>
<th>Source (Group x Level)</th>
<th>df, df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Recognition</td>
<td>2, 145</td>
<td>0.77</td>
<td>0.4654</td>
</tr>
<tr>
<td>Word Recognition</td>
<td>2, 145</td>
<td>0.14</td>
<td>0.8666</td>
</tr>
<tr>
<td>Definition Supply (Strict)</td>
<td>2, 145</td>
<td>0.09</td>
<td>0.9108</td>
</tr>
<tr>
<td>Definition Supply (Lenient)</td>
<td>2, 145</td>
<td>0.07</td>
<td>0.9332</td>
</tr>
</tbody>
</table>

There was no significant interaction between the two main effects across all test types, indicating that the proficiency level difference had no effect on the group differences. In other words, regardless of the level, the results found in Question 1 remained the same—the Combination group outperformed the other two groups.

Question 4
Interaction between Level and Group for the Delayed Recall Tests

The fourth question examined the degree of interaction between proficiency level and annotation type on tasks involving delayed recall of vocabulary. In order to answer the interaction question, all the interaction effects between Level and Group for the three instruments were examined. Table 4 lists the interaction effect between the two main effects for the three delayed tests.

Table 4
Summary of Interaction between Group and Level for the Delayed Tests

<table>
<thead>
<tr>
<th>Source (Group x Level)</th>
<th>df, df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Recognition</td>
<td>2, 145</td>
<td>0.29</td>
<td>0.7486</td>
</tr>
<tr>
<td>Word Recognition</td>
<td>2, 145</td>
<td>0.03</td>
<td>0.9676</td>
</tr>
<tr>
<td>Definition Supply (Strict)</td>
<td>2, 145</td>
<td>0.85</td>
<td>0.4282</td>
</tr>
<tr>
<td>Definition Supply (Lenient)</td>
<td>2, 145</td>
<td>1.07</td>
<td>0.3441</td>
</tr>
</tbody>
</table>

There was no significant interaction between Level and Group across the delayed tests. In spite of level, therefore, the group difference pattern remained unchanged and, thus, identical to that presented for Question 2.

Question 5
Change in Recall over Time

The last question examined whether there was a statistically significant change in learners’ vocabulary recall over time (between the immediate and delayed tests) according to the annotation type. In order to answer this question, the change...
in recall for the Picture Recognition, Word Recognition, and Definition Supply tests were compiled (see Figure 2).

Across all tests, there was a drop in scores from the immediate test to the delayed test, and the rate of decline appeared to be very similar. There seemed to be some interaction between the time of the test (immediate vs. delayed) and the group for the Definition Supply tests. A summary of all the interaction effects between time and group is given in Table 5.

Table 5
Interaction Effect between Time of Test and Group

<table>
<thead>
<tr>
<th>Source (Time x Group)</th>
<th>df</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picture Recognition</td>
<td>2, 145</td>
<td>0.09</td>
<td>0.9170</td>
<td></td>
</tr>
<tr>
<td>Word Recognition</td>
<td>2, 145</td>
<td>0.78</td>
<td>0.4611</td>
<td></td>
</tr>
<tr>
<td>Definition Supply (Strict)</td>
<td>2, 145</td>
<td>3.13</td>
<td>0.0469</td>
<td></td>
</tr>
<tr>
<td>Definition Supply (Lenient)</td>
<td>2, 145</td>
<td>2.66</td>
<td>0.0736</td>
<td></td>
</tr>
</tbody>
</table>

As seen in Table 5, the interaction effect for the Definition Supply test evaluated under the strict scoring conditions was significant ($p < .05$), while the effect for scores obtained under the lenient conditions was not. There was no significant interaction between time and group for either the Picture or Word Recognition tests. These results indicate that the time of test did not alter the relative effect of a given annotation type.
DISCUSSION

Effect of Annotation Type on Immediate Vocabulary Retention

Considering students’ pre-existing knowledge of the target words (average of .29 words as measured by the pretest), the students did retain many words incidentally as they read the text for comprehension. The average gain for the Definition Supply test (strict) was 1.4 (10%) out of 14 target words, and that for the Definition Supply test (lenient) was 3.0 (21.4%). This retention rate was comparable to rates suggested by previous studies such as Coady (1993) in which the rate ranged from 5 to 15% and Knight (1994) with a rate of 5 to 21%. The average gains for the recognition tasks were higher than the production task. The students scored an average of 6.9 (49%) on the Picture Recognition test and 6.7 (48%) on the Word Recognition test. These figures were slightly higher than 41% for the select-definition test reported in Knight’s (1994) study but lower than those found in a 1999 study by Kost, et al. in which the rates of retention were 52% and 62% for the Picture and Word Recognition tasks, respectively.

With respect to the effectiveness of annotation type, Table 1 shows that the Combination group consistently outperformed the other two groups across all measures. The Combination group did significantly better than the Text-only group for the Picture Recognition test: the effect size was .59, a medium effect according to Cohen (1988). The Combination group also outperformed the Picture-only group significantly for the Definition Supply test (strict): the effect size was .61 (a medium effect).

An interesting finding was the relative effectiveness of the Picture-only annotation. The Picture-only group performed as well as the Combination group on the Picture Recognition task, but the Text-only had significantly lower scores. This result seems logical in that the Picture-only group was exposed to the pictorial cues, even though the pictures in the test were different from those the students saw in the glosses, but the Text-only group was not exposed to such cues. In light of this finding, one would expect a similar pattern for the Word Recognition test. It is reasonable to assume that the Text-only group would do as well as the Combination group on the Word Recognition test, but the Picture-only group would probably score lower since they were not exposed to the textual cues. The results, however, failed to show the advantage of the textual cues on the Word Recognition task. Moreover, the Picture-only group scored slightly higher than the Text-only group. Similarly, trends in the analysis seem to indicate that the Picture-only group outperformed the Text-only group for the Definition Supply test. This situation suggests that the picture cues were as effective as (or even slightly better than) the textual cues for the immediate retention of word meanings. However, where the detailed or fuller meanings of the words are concerned, the Text-only group did better than the Picture-only group, as evidenced by the results of the Definition Supply test (strict). In general, the pictures were not able to convey the full meanings of the words as effectively as the textual cues.
The Effect of Text and Picture Annotation on Vocabulary Retention

Other than the aforementioned slight gain over the Picture-only group, the Text-only group scored the lowest of the three groups on all other measures. Since the conventional form of glosses is textual, the results of the study encourage the use of pictures as alternatives or as accompaniments to textual cues.

As seen in Table 3, there was no interaction between Group and Level in all three of the immediate tests, that is to say, the proficiency level difference did not have any effect on the group differences. Regardless of whether it was the beginning level or the intermediate level, the group differences maintained the same overall group pattern as that shown for Question 1.

Effect of Annotation Type on Delayed Vocabulary Retention

The figures in Table 2 indicate that after two weeks the students still remembered a good proportion of the words they had previously encountered. Again, the students did much better on the recognition tests than they did on the Definition Supply test. On the delayed test, the average score on the Picture Recognition test was 5.9 (42%) and that for the Word Recognition test was 5.37 (38%). The students, according to the results of the Definition Supply test, remembered 0.88 (6%) under the strict and 2 (14%) under the lenient conditions. These figures were slightly better than those found in Knight’s (1994) study: the study reported 35% for the delayed select-definition test and 11% for the delayed Definition Supply test.

Overall, the group differences seemed smaller. Although there were significant differences among the groups for the Picture Recognition test, the differences were not significant for the other tests. The Combination group was most effective among the three groups for the delayed tests. The Combination group outperformed the other two groups on all three measures, except for the Definition Supply test (lenient). The Picture-only group again performed better on the delayed tests than the Text-only group. On the Definition Supply test (lenient), the Picture-only group even outperformed the Combination group, however, on the Definition Supply test (strict), they scored only slightly higher than the Text-only group. For the Picture Recognition test, the Picture-only group did as well as the Combination group. For the Word Recognition test, they did better than the Text-only group. The Text-only group scored the lowest for all of the delayed tests indicating the consistent disadvantage of the textual cues compared to the picture and the combination cues. However, the differences among the groups for the delayed tests were smaller than those of the immediate measures.

Question 4 examined the degree of interaction between level and group for the delayed tests. Table 4 shows no significant interaction between the two factors in any of the tests. This result suggests that the proficiency level difference did not have any influence on the effectiveness of a particular type of annotation for the delayed tests. Both the beginning and the intermediate level students benefited most from the Combination group on the delayed tests.
Change in Recall Over Time

As seen in Figure 2, there was a drop between the immediate and the delayed test scores. The rate of decline was similar among the three groups for the Picture Recognition test, but, for the other tests, the Combination and the Text-only groups declined similarly, while the Picture-only group displayed a slightly smaller decline than other groups particularly for the Definition Supply tests.

Table 6 shows that the recognition tasks yielded a high recall rate, while the production tasks did not.

Table 6 The Retention Rate by Group on the Tests

<table>
<thead>
<tr>
<th>Group</th>
<th>Picture Recognition</th>
<th>Word Recognition</th>
<th>Definition Supply (Strict)</th>
<th>Definition Supply (Lenient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>87%</td>
<td>80%</td>
<td>61%</td>
<td>44%</td>
</tr>
<tr>
<td>Picture-only</td>
<td>88%</td>
<td>85%</td>
<td>85%</td>
<td>74%</td>
</tr>
<tr>
<td>Text-only</td>
<td>75%</td>
<td>75%</td>
<td>49%</td>
<td>64%</td>
</tr>
</tbody>
</table>

Overall, the Combination group retained the vocabulary somewhat better than the Text-only group, except on the Definition Supply (lenient) test, while the Picture-only group maintained consistently high retention rates.

As indicated previously in the results for the repeated measures (see Table 5 above), there was a significant interaction between Time and Group for the Definition Supply test (strict). There were no other significant interactions found for the other tests. The interaction between time and group indicates significant changes over time among the groups. In other words, the groups changed over time differently from one another. However, the contrasts comparing pairs of group changes over time did not show any significant differences between the Picture-only and the Combination groups or between the Picture-only and the Text-only groups for the Definition Supply (strict) and Definition Supply (lenient) test. This indicates that the rate of decline was not significantly different from one group to another.

Directions for Future Research

This study was conducted in an attempt to clarify the effectiveness of annotation type on incidental vocabulary learning. The study controlled the types of annotation and proficiency levels as important factors; however, there are many factors which could alter the effectiveness of the glosses.

The majority of this study was quantitative. Another study might offer new insights by using a more qualitative approach. For example, researchers could use a think-aloud protocol to observe what is happening and to speculate about cognitive processing when participants choose or not to look up the words.

In terms of annotation types, this study looked at single variables, Text-only and Picture only, plus a combination of these two variables. A future study could
concentrate on examining the nature of the combination more closely. For instance, what types of annotations are most effective in the combination? What kinds of pictures are suitable for the combination? What kinds of textual cues, especially when one uses the target language exclusively, are most effective for the combination?

This study concentrated on 14 target words, all of which were concrete verbs. There is a need for additional studies examining various parts of speech as well as research on ways to enhance more abstract words with pictures.

Another important factor for future research is length and repetition of exposure to the annotated vocabulary item. This experiment required only about 15 minutes for the reading activity. Longer exposure and multiple encounters with the words need to be considered in future studies.

In this study, three types of tasks were administered—a Definition Supply, Picture Recognition, and Word Recognition—in this order for both the immediate and the delayed tests. The scores of the two tasks were almost identical for the immediate (M = 6.9 for the Picture and M = 6.7 for the Word Recognition) and for the delayed tests (M = 5.9 for the Picture and M = 5.4 for the Word Recognition). Considering the fairly strong relationship between the two tasks (r = .66 for the Picture and Word Recognition for the immediate test, and r = .75 for the Picture and Word Recognition for the delayed test), a future study might consider the order of the implementation of the tasks to determine whether a different order of administration of the assessment instruments results in any unique changes in scores.

CONCLUSION

Previous studies (Knight, 1994; Krashen, 1993) support the idea that learners can retain vocabulary incidentally as they read a text for comprehension. Additional research has provided evidence for enhancements in incidental vocabulary learning when students have help such as annotations for new or difficult words (Hulstijn, 1992; Hulstijn, et al., 1996; Jacobs, et al., 1994; Watanabe, 1992). The study presented here also shows that even though the rate of incidental learning might be low, the students did learn new words in this manner.

The study also supports Paivio’s (1971, 1990) Dual-coding Theory, which advances the concept that information coded both verbally (textually) and visually (pictorially) is more effective for learning than information coded singularly. The Combination group consistently, with few exceptions, outperformed the other two groups. A comparison of the two types of annotations coded singularly showed that both the textual cues and the pictorial cues had relatively equal effects on retention, with the Picture-only group occasionally outperforming the Text-only group. However, the results show that, over time, the Combination group (the dually coded) did not retain vocabulary better than the Picture-only group (singly, visually coded) or than Text-only group (singly, verbally coded). The retention rate of vocabulary among the three groups did not significantly differ. Considering the fact that the visually coded information (pictorial cues) was able to promote
better retention, this result seems puzzling. Although various explanations may be posited, the fact remains that the nature of the pictures and texts included as elements of glosses must be explored further. Given the popularity and potential of multimedia in language teaching and learning, and in the learning of vocabulary in particular, no doubt research will soon be underway that will provide answers and perhaps more questions.

REFERENCES


It’s a cold winter night. It’s midnight, and it’s very quiet. I’m still awake and studying. I have a test tomorrow. I need to read two chapters. I finish one chapter and I read the next chapter. It’s too difficult. I can’t pass the test. Shall I keep studying? Can I take the test some other time? Shall I give up? I’m pondering many things. I think my head is going to burst.

Suddenly, some noise startles me. Something shattered on the ground. I look at the window. Wait! What is that? I see a light across the street. It is from a new house. It’s strange. Mr. & Mrs. Smith are on vacation now. They asked me to rake the lawn for them while they’re gone. Nobody should be there. Oh, I see the light again.

Then, it dawns on me. Someone is burglarizing the house. I’m afraid. What do I do now? I have to call the police. I dash to the phone and call the police.

After ten minutes, the police arrive. They enter the house. As the police search the house, someone hides outside the house. The police yell, “Stop, right there!” But the man with a black mask runs into the woods near the house. Then, he tumbles down the hill in the woods. The police finally catch him. The police take off the mask. He grins first, then, starts to sob.

Two policemen come to my apartment. The first one looks very serious. He doesn’t greet me. He just asks for my name. Then, he says, “Thank you for calling us about this problem.” The other one is friendlier. He inquires about a couple of things. He wants to know when I first saw the light. He scribbles some notes.

The policemen are gone, and everything is quiet now. What a strange night! I’m glad this is over, but I am still shivering a little. So I pour some milk. This might help me. I can’t study any longer and can’t sleep right away.

I decide to read a book. I got it at a bookstore yesterday. The title is “American Short Stories.” I look at the first chapter. And I gape at the title. It says, “My Life as a Burglar” by A Man with a Black Mask.

390 words in total
APPENDIX B
Definition Supply Test

Directions Please check any of these words you know. Please put [ X ] in the box. Please write the meanings in either English or your native language.

[ ] sob

[ ] shiver

[ ] burst

[ ] grin

[ ] scribble

[ ] ponder

[ ] shatter

[ ] dash

[ ] gape

[ ] burglarize

[ ] pour

[ ] rake

[ ] startle

[ ] tumble
APPENDIX C

Definition Supply Section Evaluation Criteria

1) **Two Points (Strict) scoring:** Students get two points when they demonstrate a clear understanding of the meaning of the word in a specific context.

2) **One Point (Lenient) scoring:** Students get a point when they demonstrate a basic understanding (the gist) of the meaning of the word.

3) **Zero Point:** Students get no point when they do not show any sign of understanding of the meaning of the word or when they do not provide any answers.

4) The model definitions are listed first, followed by other examples of correct answers.

5) Grammar and spelling are not the determining criteria for grading.

<table>
<thead>
<tr>
<th>Word</th>
<th>Two Points</th>
<th>One Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. burglarize</td>
<td>to enter (a place) without permission and steal, to rob a house,</td>
<td>to break into, to steal, to rob thief, burglar</td>
</tr>
<tr>
<td>2. burst</td>
<td>to break suddenly (by pressure from inside), to explode</td>
<td>to break, explosion</td>
</tr>
<tr>
<td>3. dash</td>
<td>to run quickly (fast, suddenly), to hurry</td>
<td>to run</td>
</tr>
<tr>
<td>4. gape</td>
<td>to look in great surprise, especially with the mouth open, to open one's mouth in surprise, to look at something in surprise</td>
<td>to surprise, to be surprised, to open one's mouth</td>
</tr>
<tr>
<td>5. grin</td>
<td>to make a very wide smile, big (wide, huge, funny) smile</td>
<td>to smile, to laugh, smile</td>
</tr>
<tr>
<td>6. ponder</td>
<td>to think about something, to think deeply, to consider, to evaluate, to consider</td>
<td>to think</td>
</tr>
<tr>
<td>7. pour</td>
<td>to make a liquid flow from one container to another, to put some drink in a container</td>
<td>to serve, to make (have) a drink</td>
</tr>
<tr>
<td>8. rake</td>
<td>to collect leaves or grass with a special garden tool</td>
<td>to take care of the garden, to clean the yard, to sweep, a special gardening tool</td>
</tr>
<tr>
<td>9. scribble</td>
<td>to write carelessly (quickly), to write in a hurry</td>
<td>to write, to take notes</td>
</tr>
<tr>
<td>10. shatter</td>
<td>to break something suddenly into small pieces</td>
<td>to break, something fall and break</td>
</tr>
<tr>
<td>11. shiver</td>
<td>to shake, as from cold or fear</td>
<td>to shake, to feel cold (chilly), to have cold</td>
</tr>
<tr>
<td>12. sob</td>
<td>to cry loudly (taking in deep breaths), to cry without control</td>
<td>to cry</td>
</tr>
<tr>
<td>13. startle</td>
<td>to jump because of a sudden surprise, to jump suddenly, to be surprised suddenly</td>
<td>to surprise, to be surprised, to jump, to get scared</td>
</tr>
<tr>
<td>14. tumble</td>
<td>to fall or roll over suddenly</td>
<td>to fall, to fall down, to roll down, to go down</td>
</tr>
</tbody>
</table>
### APPENDIX D

Picture Recognition Test

**Directions:** What does each English word mean? Please choose one matching picture. Please put [X] in the box.

<table>
<thead>
<tr>
<th></th>
<th>![Image]</th>
<th>![Image]</th>
<th>![Image]</th>
<th>![Image]</th>
<th>![Image]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. startle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. rake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. grin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. tumble</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. scribble</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. burst</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. pour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8. ponder
9. shatter
10. dash
11. gape
12. shiver
13. burglarize
14. sob
APPENDIX E

Word Recognition Test

Directions: Please match the English word with the correct meaning. Put [ X ] in the box.

1. ponder  [ ] To study for a test
            [ ] To think very carefully
            [ ] To read something
            [ ] To break open suddenly

2. shatter  [ ] To surprise someone
            [ ] To fall suddenly
            [ ] To break something into pieces
            [ ] To look outside

3. grin     [ ] To cry loudly
            [ ] To request information
            [ ] To hide behind something
            [ ] To have a big smile

4. gape     [ ] To open the mouth wide in surprise
            [ ] To get very angry
            [ ] To laugh loudly
            [ ] To shake from cold or fear

5. scribble [ ] To ask some questions
            [ ] To look at quickly
            [ ] To write carelessly
            [ ] To read carefully

6. burglarize [ ] To move or act very fast; to hurry
               [ ] To clean something very quickly
               [ ] To enter a building and steal something
               [ ] To say something loudly

7. pour     [ ] To move very fast
            [ ] To open the mouth with surprise
            [ ] To read something
            [ ] To flow from one container to another
8. startle
   [ ] To break something into pieces
   [ ] To make a loud noise
   [ ] To become very quiet
   [ ] To surprise or scare someone

9. tumble
   [ ] To say something loudly
   [ ] To hide outside
   [ ] To run away
   [ ] To fall quickly

10. burst
    [ ] To break open suddenly
    [ ] To have a headache
    [ ] To think about many things
    [ ] To feel very cold

11. rake
    [ ] To cut the grass
    [ ] To gather leaves with a special tool
    [ ] To take care of something
    [ ] To enter a building and steal something

12. shiver
    [ ] To write quickly
    [ ] To become very happy
    [ ] To shake from cold or fear
    [ ] To become very angry

13. dash
    [ ] To run very quickly
    [ ] To call someone very quickly
    [ ] To talk with someone quickly
    [ ] To become clear quickly

14. sob
    [ ] To smile very big
    [ ] To feel very sorry
    [ ] To cry without control
    [ ] To run away
AUTHORS’ BIODATA

Makoto Yoshii is Associate Professor in the Modern Communication Department at Baiko Gakuin University in Yamaguchi, Japan. He graduated from the University of South Florida with a Ph.D. in Second Language Acquisition/Teaching & Instructional Technology. He is interested in vocabulary learning/teaching in CALL.

Jeffra Flaitz is Associate Professor of Linguistics at the University of South Florida (USF) and serves as Director of USF’s English Language Institute. Dr. Flaitz is author of The ideology of English: French perceptions of English as a world language, editor of Understanding your international students: A cultural, educational, and linguistic guide to 18 countries to be published by the University of Michigan, and author of several articles on teacher preparation, language learning strategies, and language attitudes. A former Fulbright senior scholar, Dr. Flaitz has taught and carried out research in Colombia and has conducted numerous teaching training workshops in Latin America, Asia, and Europe.

AUTHORS’ ADDRESSES

Makoto Yoshii
365 Yoshimi Myooji-machi
Shimonoseki, Yamaguchi 759-6534
Japan
Phone: 0832/86-2221
Fax: 0832/86-7149
Email: yoshii@baiko.ac.jp

Jeffra Flaitz, Director
English Language Program
University of South Florida
4202 E. Fowler Avenue
Tampa, FL 33620
Phone: 813/974-3635
Fax: 813/974-2769
Email: flaitz@chumal1.cas.usf.edu