Effects of Gloss Types on Vocabulary Learning through Reading: Comparison of Single Translation and Multiple-choice Gloss Types

MAKOTO YOSHII
Prefectural University of Kumamoto

ABSTRACT
Glossing is a well-known technique for enhancement of incidental vocabulary learning. This study examines which gloss type, single translation or multiple-choice, is more effective for vocabulary learning through reading. Japanese university students read a text for comprehension on computers. The students were randomly divided into two groups: single translation gloss and multiple-choice gloss groups. There were 15 target words in the text. Glosses were provided next to the target words in the form of blank boxes with pull-down menus. The single translation gloss group checked the meanings of the target words by simply clicking the arrow and locating one definition in their L1, Japanese. The multiple-choice gloss group checked the meanings of the words in the same way; however, the students in the multiple-choice gloss group had to choose the correct definition out of four options. The participants took four vocabulary tests: 1) pretest; 2) immediate test; 3) posttest one week later; and 4) delayed posttest one month later. The tests were all multiple-choice tests except for the pretest, which employed a meaning-supply test. The results showed a consistent pattern of single translation gloss type being more effective than multiple-choice type. This was contrary to previous findings. Possible reasons and implications of the study are discussed.

KEYWORDS
Vocabulary, Glossing, Incidental Learning, Multiple-choice Gloses, Feedback

BACKGROUND
Vocabulary learning is essential for acquiring a language (Folse, 2004). Reading is a major source of incidental vocabulary acquisition (Krashen, 1985; Nagy, 1997).
Indeed, research indicates that L1 speakers learn around 15% of the unknown words they encounter incidentally through reading (Swanborn & De Glopper, 1999). L2 learners can also pick up words while reading, but the process is much less efficient (Horst, Cobb, & Meara, 1998). L2 word learning through reading often suffers from very small gains, and the process is slow and often unsuccessful. Numerous studies have recently examined the effectiveness of several techniques to enhance vocabulary learning through reading, and glossing is one of the prominent techniques (Hulstijn, 1992; Jacobs, Dufon, & Hong, 1994; Rott, 2005; Rott & Williams, 2003; Rott, Williams, & Cameron, 2002; Watanabe, 1997). Among many types of glosses, this paper pays particular attention to single-translation glosses and multiple-choice glosses.

LITERATURE REVIEW

Previous Research on Glosses

Glossing provides definitions or explanations of obscure words in the margins of a text. Glosses direct readers’ attention to unfamiliar words and encourage the processing of the meanings of the words. Studies have examined the effectiveness of glosses for incidental L2 vocabulary learning for printed materials and found that glosses enhanced vocabulary learning (Hulstijn, 1992; Hulstijn, Hollander, & Greidanus, 1996; Jacobs et al., 1994; Knight, 1994; Paribakht & Wesche, 1996; Watanabe, 1997). With advances in technology, glosses expanded beyond simple texts into multimedia glosses including texts, graphics, videos, and audios (Al-Seghayer, 2001; Chun & Plass, 1996). The previous studies on multimedia glosses found that L2 learners learned unknown vocabulary items better when they looked up a combination of pictures and text definitions than when they looked up definitions alone (Abraham 2007; Chun & Plass, 1996; Plass, Chun, Mayer, & Leutner, 1998; Yeh and Wang, 2003; Yoshii, 2006; Yoshii & Flaitz, 2002). Even though multimedia has brought a variety of types of glosses, L2 learners often prefer textual information among the different types (Davis & Lyman-Hager, 1997; Erçetin, 2003; Laufer & Hill, 2000; Laufer & Kimmel, 1997; Lomicka 1998; Pardo-Ballester & Rodríguez, 2010). Text-based glosses in CALL, therefore, are still important and deserving of further investigation to find ways to utilize them.

Studies on Textual Glosses

Studies on textual glosses have dealt with languages, e.g., L1 or L2, and formats of glosses, e.g., basic dictionary format or sentence-level translation. Regarding the languages, researchers have examined in which language, L1 or L2, glosses should be written and the results indicate that the effectiveness of the glosses in L1 or L2 may depend on learners’ proficiency levels. High-proficiency learners can benefit from L2 glosses while low-proficiency learners can gain much from L1 glosses (Hulstijn et al., 1996; Jacobs et al., 1994; Yoshii, 2006).

Studies have examined the effectiveness of the glossing formats of textual types, e.g., basic dictionary format or sentence-level translation (Gettys, Imhof, and Kautz, 2001; Grace, 1998). Gettys et al. (2001) compared a basic dictionary
form with a sentence-level translation equivalent. The basic dictionary form used the definition straight from a dictionary. The sentence-level format came out of the context and used a translation that fit in the sentence. The researchers found that the former group retained vocabulary better than the latter group did. The researchers argued that the longer time to read the text with the dictionary form resulted in deeper processing. Grace (1998) investigated the effectiveness of sentence-level translations, that is, the translations of the whole sentences in which the words appeared. The researcher argued that sentence-level translations would make learners process lexical information deeper since they have to infer and derive the meanings of the words from the clues in the translations. The researcher found that the learners with sentence-level translations learned better than those without the translations both for short-term and long-term learning. The results tell us the importance of both deep processing and verifications of meaning.

**Theoretical Underpinnings of Multiple-choice Glosses**

Studies have also examined the effectiveness of single translation and multiple-choice glosses. Single translation glosses are ordinary, conventional glosses with one definition or one explanation of a word; on the other hand, multiple-choice glosses have multiple definition options—typically one correct definition of the word in question and three definitions of other words—and learners have to think about the meaning of the word and choose the best one that would fit the context where the word appears. The theoretical underpinnings of many studies on multiple-choice glosses are based on levels of processing depth theory proposed by Craik and Lockhart (1972). The theory states that the chances of storing new information in long-term memory depend on how deeply it is processed besides attending to it during its occurrence and rehearsing it after its occurrence. Hulstijn (1992) used the term “mental effort” to explain the depth theory, stating that inferences and hypothesis-testing of word meaning would lead to better word retention. Laufer & Hulstijn (2001) applied the concept of depth of processing to the SLA field and introduced the Involvement Load Hypothesis. The theory states that learners can learn words better when they are highly involved in lexical information processes. The theory consists of three components: a motivational one (‘need’) and two cognitive ones (‘search’ and ‘evaluation’). ‘Need’ means that one has to learn a word because someone else tells him or her to do so, or one does so out of curiosity or necessity. ‘Search’ is to find the meaning of an unknown word. Finally, ‘evaluation’ is defined as “a comparison of a given word with other words, a comparison of a specific meaning of a word with its other meanings, or combining the word with others in order to assess whether a word (i.e., a form-meaning pair) does or does not fit its context” (p. 14). The presence or absence of these components determines the degree of involvement of tasks; thus, in turn, the involvement load affects vocabulary retention.

Going back to the topic of multiple-choice glosses, Laufer & Hulstijn (2001) used them and hypothesized that they would be better than single translation glosses because multiple-choice glosses would induce more involvement of
learners than single translation glosses would. The single translation gloss group read a text using glosses with single translations and did not have to search for the meaning of the words or engage in the evaluation process (+Need, -Search, -Evaluation). On the other hand, the multiple-choice gloss group read a text with multiple-choice glosses and had to search for the meaning of the words and evaluate which meaning would best fit among different options (+Need, +Search, +Evaluation). The presence of all three components in multiple-choice glosses in comparison to single translation glosses with one component led to the hypothesis that multiple-choice glosses would be better than single translation glosses for word retention, although this hypothesis has not been supported fully by single translation gloss and multiple-choice gloss studies.

**Studies on the Comparison of Single Translation Glosses and Multiple-choice Glosses**

Rott and others have conducted a series of studies to investigate qualitatively the effectiveness of multiple-choice glosses using think-aloud protocols (Rott & Williams, 2003; Rott et al., 2002). The studies reported the effectiveness of multiple-choice glosses over no gloss conditions. However, comparisons between single translation glosses and multiple-choice glosses have brought mixed results. Some studies did not find any significant differences between the two types for incidental vocabulary learning (Hulstijn, 1992; Miyasako, 2002; Watanabe, 1997), while others indicated that multiple-choice glosses were more effective than single translation glosses (Nagata, 1999; Rott, 2005).

Glosses usually contain an L1 translation or L2 synonym in the margin of the text. Multiple-choice glosses display multiple L1 word meanings and readers have to choose which meaning would fit best. This supposedly reduces the likelihood of making an incorrect form-meaning connection compared with reading a text without such glosses. Multiple-choice glosses also engage the reader in the cognitive process of evaluation, which is claimed to foster word retention (Laufer & Hulstijn, 2001). Evaluation here means to check to see if a correct meaning was chosen. That is, the provision of multiple-choice glosses in the margin of a text appears to be an intervention that triggers a higher involvement with a word than “normal” glosses. In a series of studies, Hulstijn (1992) observed that multiple-choice glosses resulted in more word gain than L2 synonym glosses, but also that L1 glosses resulted in more word gain than multiple-choice glosses. In addition, the study showed that a significant number of participants in the multiple-choice condition chose a wrong meaning. This system is still susceptible to wrong choices by learners unless they have a chance to confirm the choices. Hulstijn suggested that four alternatives might be too many.

Watanabe (1997) compared four formats for presenting an English reading text to Japanese students: (1) appositives (inserting an L2 definition immediately after each word in the text), (2) single marginal glosses (providing such an L2 definition in the margin of the text), (3) multiple-choice marginal glosses (providing two L2 definitions in the margin of the text and having to choose one definition),
and (4) control (text only). The study found that both single gloss and multiple-choice gloss groups significantly outperformed the appositive and the text-only groups on the vocabulary posttests. Concerning the comparison of two marginal glosses, the single gloss group performed better than the multiple-choice gloss group; however, the difference was not statistically significant. This experiment was conducted with printed materials, and the multiple-choice group did not have any means to check to see which L2 definition was actually correct. The lack of confirmation of their choices might have been connected to the ineffectiveness of the multiple-choice glosses.

Miyasako (2002) examined four types of glosses along with text-only and control (no reading). The four types consisted of (1) multiple-choice glosses in L2, (2) multiple-choice glosses in L1, (3) single gloss in L2, and (4) single gloss in L1. This was a paper-based study and the glosses appeared in the margin of the text. The multiple-choice glosses contained two definitions either in L1 or L2 and the students were told to select one definition. Immediate and delayed (18 days later) vocabulary tests were conducted. The study found that the L2 multiple-choice gloss group scored higher than other gloss groups, but significant group differences did not emerge at the immediate or at the delayed tests. L2 gloss groups performed better than L1 groups at the immediate test, but no difference appeared in the delayed tests. The effect of gloss types had a relationship with English ability: L2 glossing was more effective for higher-ability learners, while L1 glossing was effective for lower-ability learners.

Nagata (1999) also compared two types of glosses in a Japanese reading program. American university students read an L2 text in Japanese on computers in two formats: (1) single glosses (providing an L1 translation for each word) and (2) multiple-choice glosses (providing two L1 translations). The gloss content appeared on the side of the screen as the learners clicked on the words in the text. The multiple-choice gloss group selected an L1 translation and was able to check the correct answer through feedback given immediately. The study showed that the multiple-choice group performed significantly better than the single-gloss group on the immediate posttest. The researcher suggested the effectiveness of the multiple-choice gloss in this study came from the fact that it provided students with immediate feedback on their selections as well as encouraging deeper lexical processing. However, the advantage of the multiple-choice group did not last more than one month, and a significant difference did not emerge on the delayed posttest.

Rott (2005) compared the effectiveness of multiple-choice glosses and single translation glosses through think-aloud procedures. She gathered detailed descriptions of how 10 learners processed the glosses as they read a text for comprehension. The results of four target words measured by VKS (vocabulary knowledge scale) revealed that multiple-choice glosses were more effective than single translation glosses for strengthening form-meaning connections.

As seen above, the number of studies is still limited and the results are still not conclusive. The effectiveness of single translation glosses and multiple-choice glosses seems to be related to how deeply one can process the lexical information
presented by glosses and whether one can have immediate feedback on multiple-choice glosses. Watanabe (1997) and Miyasako (2002) conducted paper-based studies and did not give any immediate feedback on multiple-choice glosses. Consequently, no difference was found between single translation glosses and multiple-choice glosses. Nagata (1999), on the other hand, conducted the study on computers, providing immediate feedback to the multiple-choice gloss group, resulting in better performance by the multiple-choice gloss group over the single translation gloss group. However, the difference in effectiveness did not last more than one month. It is important to compare single translation glosses and multiple-choice glosses in the computerized condition where immediate feedback is available and to see whether it has any long-term effects.

THE STUDY

Purpose of the Study & Research Questions

The purpose of the current study is to compare the effectiveness of single translation and multiple-choice glosses for incidental vocabulary learning. The study aims to examine the short-term and longer-term effects of each gloss type in the computerized condition where immediate feedback is given upon the request of a learner.¹ The research questions are as follows.

Reading a text for comprehension purposes,
1. which gloss type, single translation gloss type or multiple-choice gloss type, is more effective for short-term vocabulary retention?
2. which gloss type, single translation gloss type or multiple-choice gloss type, is more effective for long-term vocabulary retention?
3. are there any interaction effects between the gloss types and time of tests for vocabulary retention?

Method

Participants

The participants were 41 university students in Japan who were majoring in English. They were freshmen and had had at least six years of learning the language at their junior high and high schools. They consisted of 11 male and 30 female students and their average age was 19 years. Their English proficiency was considered as low-intermediate as measured by a standardized test, the TOEIC®. The average score was 497 points ranging from 250 to 675, which is equivalent to 470 points in TOEFL score, roughly speaking, ranging from 383 to 530. As seen in the lookup behavior section later, the final analysis involved data from 35 students consisting of 10 male and 25 female students. The reason for the attrition will be explained later.

Reading material

The text used was taken from a textbook and was about Jean-Jacques Rousseau, symbolizing how one man can change the world through his ideas and writings.
The text contained about 300 words consisting mostly of Level 1 & 2 words (87%) in the General Service List (West, 1953). The list introduces the first 1,000 (Level 1) and the second 1,000 (Level 2) most frequently used English words. Readability measured by Flesch-Kincaid (Flesch, 1948) points to grade level 12, and the reading ease score was 41, showing that the text was considered intermediate to high-intermediate for L2 learners. The study was conducted using Moodle, a class management system widely used throughout the world. The system was in use from the beginning of the semester. Therefore, the students were already familiar with the system and reading texts on computers. In order to create a text with glosses embedded in the text, the researcher used the cloze-making function in the Moodle installation. Using this function, one can create a pull-down menu of glosses immediately after a word. Figure 1 displays a screenshot of the text which appeared on the Moodle page.

**Figure 1. Screen shot of the text on Moodle**

![Screenshot of text](image)

Target words

There were 15 target words: one phrase “live off” from Level 1, one (“overcome”) from Level 2, two words (“erroneous” and “inherent”) from the Academic Word List (Coxhead, 2000). A further 11 words (“blueprint”, “despondent”, “famine”, “indifferent”, “ingrain”, “nobility”, “oppressed”, “peasantry”, “placid”, “sages”, and “weary”) came from outside the lists, meaning that they were infrequent and
difficult words.

**Procedures**

There were 41 participants in the study. They were all university students in Japan. They were first stratified by their English proficiency measured by a general proficiency test. Then, they were randomly assigned to two groups, single translation gloss (SG) group or multiple-choice gloss (MG) group. Before the reading activity, they took a pretest of target words. It was a definition-supply test and learners had to write the meaning of a word either in L1 or L2. Then, they read a text in Moodle as seen in Figure 1. The students read a text in either SG or MG condition for comprehension purposes.

The students were instructed to guess the meanings of the words. If they were not sure of the meaning or if they did not know the meaning, they could click on the arrow to see a definition. Upon each click, the SG group could see a definition of the word in L1. The MG group saw four definitions in L1. The choices included one correct meaning, two additional meanings that would make sense in the context, and one meaning from other target words. The MG group had to choose one definition out of four as the most appropriate meaning for the word. An example of SG is seen in Figure 2, and an example of MG is seen in Figure 3.

**Figure 2. Example of single gloss**
Figure 3. Example of multiple gloss

Note. The translations of the Japanese in the distractors for the word, “famine,” are as follows from top to bottom: 飢餓 = famine; 豊作 = farming; 著名 = famous; and 家系 = family tree.

There were five multiple-choice reading comprehension questions at the end of the text. After reading and working on the comprehension questions, they clicked on the Submission button. Then, a feedback page appeared as seen in Figure 4. The SG group saw this feedback page reminding them of which words they clicked and which words they did not. The boxes right after the words the learners clicked were shaded with the definitions, and those that the learners did not click remained as blank boxes. The MG group saw the feedback page as Figure 5. They saw which words they got correct and which ones incorrect. When they clicked on the arrow on the incorrect ones, a pull-down menu appeared, indicating what the correct answer was. Learners’ lookup behaviors were recorded in log files.
Figure 4. Example of feedback page for the SG

Many of us who despair about the world's problems of overpopulation, pollution, global warming and famine feel helpless to change things. While we care about these problems, and may even feel despondent by them, there seems to be nothing we can do about them.

But for inspiration we should study the lesson of one gentle soul who lived more than 200 years ago. The lesson of this man is not how to overcome such specific global problems, but rather that one person—even alone—can make a difference.

This man, who was born in Switzerland and spoke French, wrote books about human rights and social equality. He believed that the greatest problems in the world were ingrained in the erroneous belief that money and political power made one person more worthy than another. He observed that 'civilization' destroyed the inherent value of the individual.

In his day, a small percentage of the population was living off the blood and sweat of a vast mass of oppressed and soul weary peasantry. The nobility did not work for a living but enjoyed luxury beyond belief. At the head of this system were a self-righteous king and an indifferent queen.

His writings became the blueprint for both the French and American revolutions, although he himself would never have approved of violence. The later constitutions of these nations were also based on his beliefs.

In setting forth his views that all people are created equal and that everyone should be given the opportunity to develop themselves to their fullest potential, this placed gentle man changed the foundation of the western social and political system. He introduced the concept of human rights, which we now take for granted.

This man will go down in history as one of the world's greatest sages. His name is Jean-Jacques Rousseau.

Figure 5. Example of feedback page for the MG

Many of us who despair about the world's problems of overpopulation, pollution, global warming and famine feel helpless to change things. While we care about these problems, and may even feel despondent by them, there seems to be nothing we can do about them.

But for inspiration we should study the lesson of one gentle soul who lived more than 200 years ago. The lesson of this man is not how to overcome such specific global problems, but rather that one person—even alone—can make a difference.

This man, who was born in Switzerland and spoke French, wrote books about human rights and social equality. He believed that the greatest problems in the world were ingrained in the erroneous belief that money and political power made one person more worthy than another. He observed that 'civilization' destroyed the inherent value of the individual.

In his day, a small percentage of the population was living off the blood and sweat of a vast mass of oppressed and soul weary peasantry. The nobility did not work for a living but enjoyed luxury beyond belief. At the head of this system were a self-righteous king and an indifferent queen.

His writings became the blueprint for both the French and American revolutions, although he himself would never have approved of violence. The later constitutions of these nations were also based on his beliefs.

In setting forth his views that all people are created equal and that everyone should be given the opportunity to develop themselves to their fullest potential, this placed gentle man changed the foundation of the western social and political system. He introduced the concept of human rights, which we now take for granted.

This man will go down in history as one of the world's greatest sages. His name is Jean-Jacques Rousseau.

Note. The translations of the Japanese for the wrong answers are as follows: 命を絶つ = to commit suicide; 貧困の = poor; 移動 = transfer; 平等 = equal; 時代 = age
After the reading activity, they were unexpectedly given a vocabulary test (immediate test). It was a multiple-choice test, and the learners had to select the best definition for each word out of four options, which were the same distractors in the reading activity. The definitions were written in L1. One week later, the learners took a posttest (delayed test 1). Then, one month later, they took another delayed posttest (delayed test 2). All three tests consisted of the same questions; however, the order of the questions and the order of the choices for each question were randomized for each test and for each learner.

**Data Analyses**

In order to compare the effectiveness of the gloss types at each test, the data were originally intended to be analyzed using two sample *t*-tests for immediate test, delayed test 1 (one week later), and delayed test 2 (one month later). However, as seen in the analyses of lookup behaviors, the number of the participants in SG for the final analyses dwindled from the original 20 students to 14 students. This raised a concern for the equality of variances of the two groups, since the two groups now differed in number (SG n=14, MG n=21), and the number of the participants for each group was rather small. One way to deal with the problem was to conduct an *F*-test to compare two variances. If the two were not significantly different, a two-sample *t*-test could be used. If they differed, non-parametric analysis would be used. However, when a *t*-test or non-parametric test is done after *F*-test, type I errors could be multiplied. Therefore, in this study, it was decided to use the Welch two-sample *t*-test from the start. The Welch test makes it possible to examine the differences of two groups with different variances. Effect sizes were also calculated using Cohen’s *d*. While *p*-values often depend on the sample size, effect sizes are independent of the size. The effect sizes using Cohen’s *d* were interpreted as follows: *d* > .20 = small, *d* > .50 = medium, and *d* > .80 = large (Cohen, 1992).

The data were also analyzed by a mixed design repeated measures of 2 (gloss types) X 4 (tests) ANOVA to see the overall effect of the gloss types over time. The study had one between-subjects factor, Gloss types (single translation glosses or multiple-choice glosses), and had one within-subjects factor, Tests (pretest, immediate, delayed 1, & delayed 2). The effects sizes were sought after using η². The effects sizes with η² were interpreted as small with η² > .01, medium with η² > .06, large with η² > .14 (Cohen, 1992).

The dependent variables were vocabulary test scores measured by definition-supply test for the pretest and multiple-choice tests given immediately, one week later, and one month later. In addition to *p*-values, effect sizes were also noted. The level of statistical significance was set at .05 as an alpha level in all statistical analyses.
RESULTS

Lookup Behavior

Before analyzing the data for the research questions, the results of lookup behaviors need to be introduced first. The reading program recorded the learners’ lookup behaviors of the target words. This helped the researcher to examine if the learners really looked up the targeted words in the text, especially for the multiple-choice gloss (MG) group since they had to select the meanings of the target words. As Fischer (2007) points out, it is essential to monitor students’ behavior in CALL and to find out what students are actually doing. Without this information on learners’ actual behaviors, one runs the risk of making erroneous assumptions.

Figure 6 shows the results of the lookup behaviors of the single translation gloss (SG) group. The target words were listed horizontally from target word#1 (wd1) to target word#15 (wd15) in alphabetical order. The students were listed vertically from student#1 to Student#20. The students were arranged according to the number of lookups they made. The words that the learners looked up are in gray and those the learners did not look up are in white. As seen in the bottom part of Figure 6, six students in the Single Gloss group hardly looked up any words: one student (student#15) looked up only three words, the next student (student#16) looked up only one word, and the last four students (student#17~20) did not look up any words at all. This study is based on the assumption that the learners used the glosses while reading. Therefore, it became necessary to remove those six students. Consequently, the number of participants in the SG group decreased from 20 to 14.

Figure 6. Lookup behavior of Single gloss group
Figure 7 displays the results of the lookup behaviors of the MG group. The MG group looked up all the words available to click; however, they did not necessarily select the correct meanings, as indicated by the light gray cells. These cells indicate that the learners did look up but made incorrect choices. This has previously been found to be the main drawback for multiple-choice glosses (Hulstijn, 1992; Rott, 2005). Readers in the MG group may not be able to infer a correct word meaning and succeed in establishing correct initial form-meaning connections. In fact, Figure 7 shows the students chose 113 incorrect answers out of 315 words (15 words x 21 students). This amounts to 36%, one third, of the total words. Nonetheless, all 21 students were counted for the study since they did look up the words regardless of the correctness of their choices of the meanings.

**Figure 7. Lookup behavior of MG group**

<table>
<thead>
<tr>
<th>st1</th>
<th>wd1</th>
<th>wd2</th>
<th>wd3</th>
<th>wd4</th>
<th>wd5</th>
<th>wd6</th>
<th>wd7</th>
<th>wd8</th>
<th>wd9</th>
<th>wd10</th>
<th>wd11</th>
<th>wd12</th>
<th>wd13</th>
<th>wd14</th>
<th>wd15</th>
</tr>
</thead>
<tbody>
<tr>
<td>st2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>st21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Reading Comprehension**

The learners worked on the comprehension while or after reading the text. This was the main task for the reading, and the vocabulary test was not announced beforehand. The comprehension questions consisted of five questions. The learners had to choose the correct one out of the four options. The original data with all 41 students showed that the SG group had a mean of 3.40 (SD = 1.19) and the MG group had a mean of 3.81 (SD = 1.12), both having Max. = 5. T-test did not show any significant difference between the two groups: $t = 1.14, DF= 39, p = .26$. The effect size, measured by Cohen’s $d$, was .37, small to medium. When six students were excluded due to the lack of lookup behaviors in the SG, the mean for the group became 3.00 (SD = 1.18), and the Welch $t$-test still did not show
significant difference between the two groups: $t = 2.03, DF = 27.10, p = .05$. However, a slight difference seemed to emerge since the effect size was $d = .78$, large. This may indicate that those six students who were excluded in the latter analysis contributed to the former analysis. This, in turn, may indicate that those six students had good reading comprehension scores. If they had a good reading comprehension level, one of the reasons for them not to look up words was that they did not feel the need. They were able to guess the meanings of the words well. Some of them, of course, might have ignored the glosses altogether for one reason or another.

**Vocabulary Tests**

Pretest

Before analyzing the data for the research questions, the results of the pretest are discussed first. The pretest was conducted before the reading activity in the form of L2-L1 translation. The test was given in the Moodle installation. The students had to supply an L1 translation for each L2 word. The SG group ($n = 14$) had a mean score of 1.93 ($SD = 1.94$) while the MG group ($n = 21$) had a mean score of 1.71 ($SD = 1.31$). The maximum score was 15 points. This means that the learners’ pre-knowledge of the target words was less than two words. Welch $t$-test did not indicate a significant difference between the groups ($t = 0.36, DF = 20.84, p = .72, Cohen’s d = .16$).

Immediate Test, Delayed Test 1, and Delayed Test 2

Before analyzing the data, an overall picture of the test results is shown in the following Figures 8-10. First, the results of both groups were displayed in the order of immediate test, delayed test 1 (one week later), and the delayed test 2 (one month later). All three tests were multiple-choice tests. The contents of the tests were the same. The order of the questions and the order of the choices were randomly assigned for each learner as they logged in. As was the case for Figures 6 & 7, each Figure has target words lined horizontally from word#1 (wd1) to word#15 (wd15). The learners were listed vertically from student#1 (st1) to student#14 (st14) for the SG group and from student#1 (sd1) to student#21 (sd21) for the MG group.
Figure 8. Immediate test results of SG & MG groups

SG group

MG group
Figure 9. Delayed test 1 results of SG & MG groups

**SG group**

![Table for SG group]

**MG group**

![Table for MG group]
Figure 10. Delayed test 2 results of SG & MG groups

**SG group**

As seen in Figure 8, the learners in both groups learned most of the words. The SG group outperformed the MG in the immediate test since the white cells, indicating the incorrect answers for the multiple-choice test, are more evident than those in SG. As time passed, the students’ knowledge of the words decreased. One can see this in the gradual spread of the white cells from Figures 8 to 10. The rates for the expansion of the white cells in MG seem to be more obvious than those in the **MG group**.
Now the comparison of each test will be examined statistically between the SG group and the MG group. Figure 11 shows the results including the pretest. The inclusion of the pretest helps to capture the learning effect observed in the sudden jump of the scores at the immediate test. The cause of the sudden increase of the scores is most likely due to the learning taking place. The cause could be attributed to the differences between the test types as well. The pretest was conducted through L2–L1 translation (meaning-supply task); while the immediate test employed a multiple-choice test. The latter seems to be easier than the former; therefore, the cause of the sudden increase could partially be attributed to the nature of the multiple-choice test which tends to lead to higher scores than meaning-supply tests.

Figure 11. Comparison of SG and MG at each test
Research Question 1: Short-term effect of the gloss types

As Table 1 indicates, the results of the immediate test revealed significant differences between the two groups ($t = 2.41$, $DF = 30.29$, $p = .02$, $d = .87$). The SG was more effective than the MG in the short term, as measured by the immediate test.

Research Question 2: Long-term effect of the gloss types

The results of the posttest (Delayed 1) given one week after the treatment revealed that there was still a significant difference between the two groups ($t = 2.73$, $DF = 32.64$, $p = .01$, $d = .95$). The SG group continued to outperform the MG group even though both groups scored less than in the immediate test, indicating that memory loss has occurred in one week.

The delayed posttest (Delayed 2) was administered one month later. The results followed the previous ones, indicating the advantage of the SG over the MG ($t = 2.10$, $DF = 32.59$, $p = .04$, $d = .73$). The effect sizes were all rather large; the immediate with $d = .87$, the delayed 1 with $d = .95$, and the delayed 2 with $d = .73$. The difference between the two groups seemed to have widened a little at the delayed 1 from the immediate test, and narrowed a little from the delayed 1 to the delayed 2.

Research Question 3: Overall Changes Over Time

One can see the overall changes over time in Figure 12. In order to see the overall effect of gloss types over time, an ANOVA was conducted.
The ANOVA revealed that there was no significant interaction effect \((p = .07)\), and the effect size was small \((\eta^2 = .005)\). The significant differences were found between Gloss types, \(F(1, 33) = 5.49, p = .025, \eta^2 = .018\) and Test, \(F(3, 99) = 399.20, p = .000, \eta^2 = .80\). They show that the advantage of the SG over the MG was constant over time as seen in Figure 12. Both groups showed a remarkable increase in scores from the pretest to the immediate tests. The students learned a dozen words through reading the text with the glosses. As the lines in Figure 12 drop gradually at each posttest, some forgetting started. However, the forgetting rate is still not very steep. Both groups were able to retain the words over one month, the SG group still performing best at the last test.

**DISCUSSION**

This study examined two gloss types, single translation and multiple-choice glosses. As Figure 12 shows, scores on the vocabulary tests increased from the pretest to the immediate test. The SG and the MG groups learned about 12 words (80%) and 11 words (73%) respectively out of 15 target words. This is a confirmation of the effect of glossing for enhancing incidental vocabulary learning through reading. The rate of learning words incidentally without glossing in the L2 is discouragingly low. Horst et al. (1998) reported a pickup rate of one word out of every 12 words (8%) in early research, and in their own study, a rate of one new word out of every five words (20%). The effect of glossing in this study showed a 70 to 80% pickup rate on the immediate multiple-choice test. This is a reminder of the effectiveness of glossing as an enhancement for vocabulary learning through reading.

The results showed that the SG group consistently outperformed the MG group in all the three posttests: immediate, delayed 1 (one week later), and delayed 2 (one month later). Statistically significant differences were detected in all the
The ANOVA results also indicated that the better performance by the SG than the MG was consistent over time. Therefore, it is safe to say that the SG outperformed the MG overall.

This result was contrary to previous studies which found the MG group outperformed the SG (Nagata, 1999; Rot, 2005) or which did not find any significant differences between the two groups (Hulstijn, 1992; Miyasako, 2002; Watanabe, 1997). What are the reasons for the discrepancy? We will look at possible reasons from the perspectives of the cognitive load hypothesis and from the Involvement Load Hypothesis.

First, we will examine the results from the perspective of cognitive load and one of the key components of the Involvement Load Hypothesis, “search.” Those who support the use of MG argue that MG is better than SG since MG helps learners process meanings of words at a deeper level. While the SG group simply looks at the meaning provided (“-search”), the MG group has to think about the best choice among the options (“+search”), which brings a deeper processing in the latter process. However, it is almost impossible to find out how learners actually process the information. Do they really think about the choices and choose the one in MG? Do they simply look at the meaning in SG? Do they guess the meaning in some cases? In this study, the learners in both groups were encouraged to guess the meanings of the words first, then, if they did not know or were not sure about the words, they could look them up. The glosses were made in a way to encourage this process, giving a blank space for the gloss and the opportunity to guess. Then, when the learners felt a need to look it up, they could click on an arrow to find a meaning. Again, it is unknown what exactly each learner did when he or she clicked on each gloss. It is possible for the SG group to have engaged in a deep level of processing through guessing the meanings (“+search”) and to have clicked to check if the guessing was correct or not. This might have contributed to the advantages of the group. What exactly the learners do or process through SG is a topic for future study. As Rott and others (Rott & Williams, 2003; Rott et al., 2002) have conducted studies on the effectiveness of MG, similar studies using think-aloud protocols are needed to qualitatively examine the effectiveness of SG in the computerized texts.

Secondly, we will examine the results from the perspective of cognitive load and another key component of Involvement Load Hypothesis, “evaluation.” “Evaluation” means to see which meaning would fit a particular context in which a word occurs. It is one thing to evaluate the meaning, but it is quite another to know whether the evaluation was correct. It is essential for learners to have an opportunity to check their evaluations to see if they are correct or not. Otherwise, learners may fail to establish correct connections between meanings and forms of words at this initial stage. Some of the studies done in a paper-based environment (Hulstijn, 1992; Rot, 2005) acknowledged this drawback of MG and suggested it as one of the reasons why MG was not as effective as SG. Nagata (1999) provided MG in computerized texts and asked learners to confirm their evaluations with immediate feedback on their choices. This was a partial replication of Watanabe (1997) in a paper-based study. Nagata found that the MG was better than the SG,
contrary to the results of Watanabe's study, which found no differences between the two. The effectiveness of the MG in Nagata's study was enhanced by the confirmation process which immediately followed evaluation. In this present study, the MG group did receive feedback on their evaluation as seen in Figure 5. Then, why did the SG group still outperform the MG group? The reason may lie in the ways the feedback was presented to the MG group.

The students in the MG group could not make the best use of the MG because of the timing of the feedback, divided attention, and the processes of arriving at correct answers. First, the timing of the feedback might be important. The MG group received the feedback on the selection of the meanings only after all the work was done, including reading the text, choosing a meaning for each target word, and answering the reading comprehension questions. When they clicked the button, "Finish and send the results," the feedback page appeared as shown in Figure 5. On the other hand, the SG group saw the feedback on each word while engaging in the reading activity, not after the reading. The SG group had the advantage of receiving instant feedback on each word while reading. The MG group had to wait for the end of all the tasks. This delay in getting the feedback might have been one of the reasons for the ineffectiveness of the MG.

Secondly, the SG group was able to pay full attention to each word while the MG group had to divide their attention between several words at the feedback stage. The SG group, as mentioned above, saw instant feedback on each word during the reading. They were able to think about the word meaning in context and verify the guess (if they had guessed) with the correct meaning one by one. When learners can deal with one word at a time as in the SG group, it is likely to help them concentrate on the particular word in context and this would strengthen the initial connection between the correct meaning and the form of the word. The MG group, however, had to face all the words at once at the feedback page as shown in Figure 5. Consequently, they had to divide their attention and distribute their cognitive resources between several words. In addition, each word had four answers to choose from. This is in contrast to Nagata (1999), in which the number of choices was only two. Only two choices seemed to be too simple to promote deeper processing. On the other hand, the four choices in the present study may have been too many, preventing learners from processing the words and remembering them. In the process, the learners may forget which correct meaning is attached to which word. The difficulty of paying full attention to several words with many options at once may have been related to the inefficiency of the MG.

Thirdly, the way to arrive at the correct meaning at the feedback stage might be the reason for the ineffectiveness of the MG. Members of the SG group were able to check and confirm the meaning in one step, that is, by clicking on the arrow on the box immediately after the word. The MG group, on the other hand, had to, first, identify the incorrect choices indicated by the different colors in the feedback page seen in Figure 5. Then, if the learners wished to check the correct meaning, they had to click on the arrow to locate it in the pull-down menu. It took
two steps to arrive at the correct meaning, identifying the incorrect choices made and locating the correct one. This extra work may have inhibited some learners from actually checking and confirming the correct meanings. As a result, the MG group may not have performed as expected.

In summary, the MG group had a disadvantage during the process of “evaluation.” They were able to check and confirm their choices quickly, but still not quickly enough, unlike the SG group. The MG group had to confirm the meanings of all the words at once. Divided attention and distribution of cognitive resources might be partially related to the inefficiency of the MG. Finally, the extra step the MG group had to take to find the correct meaning might also be a partial reason for the disadvantage of the MG. Even though the feedback was given, the confirmation process was still not ideal. What if the students were able to check their choices word by word? In other words, if, rather than waiting for the end of all the treatment, while reading, the learners clicked on the choice and also checked the meaning right away, this might have brought different results.

NEED FOR FURTHER STUDY & CONCLUSIONS

Considering what has been said in the discussion, it would be desirable in a future study to devise the MG group in such a way that a learner could get the confirmation of the evaluation immediately after selecting the answer rather than waiting for all the tasks to be finished. Immediate feedback on the word choices might bring different results than those seen in this study.

Future work should include a qualitative part, for instance, implementing a think-aloud protocol, interviewing learners, giving questionnaires, etc. This study faced several issues in which qualitative information about the participants and how they processed the words would have shed light on the questions. For instance, why did some of the learners hardly look up the words at all? How did the learners use the SG? Did they guess before looking at the provided glosses? What did the learners do with the MG? How much guessing was involved when learners chose an answer? When the feedback page appeared, did the learners check their answers on the words? Did they take time to confirm their answers on the meanings of the words? In order to find answers to these questions, both quantitative and qualitative data are needed, especially think-aloud protocols would be helpful to find out what the learners experience when working with the SG or the MG (Rott, 2005; Rott & Williams, 2003; Rott et al., 2002).

In this study, the L1 was used to provide the meanings or definitions of the words. A future study might want to consider using the L2 exclusively with the SG and the MG and to see if there are any differences between the two types. Proficiency levels of the learners would make a difference and need to be included as an important factor in the study. More studies are needed to see what more advanced learners do with single or multiple-choice glosses.
ACKNOWLEDGMENT
This work was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI)(C)(23520691).

NOTES
1. In this study, “immediate” is used comparatively. When paper-based materials are used, it usually takes time for an instructor to collect them, grade them to see which choices of definitions were correct and which were not, and return them to learners. It may take from one day to one week for the learners to know the correct definitions of words depending on how often the class meets in a week. When computerized materials are used, it is immediate since learners will know the correct definitions immediately after making such a request. However, due to the design of the reading materials using Moodle’s cloze-making functions, there was a difference in the meaning of “immediacy” between SG group and MG group. The SG group was able to see the feedback immediately after clicking on each word during reading. The MG group was able to see the feedback immediately after finishing the reading. The details of the design difference can be also seen in Note #3. Even though there was such a difference, both types of feedback could still be regarded as immediate, compared to the case in paper-based materials.

2. An anonymous reviewer raised a question about the validity of the term “feedback” for the SG group since they clicked on the SG’s translation. The meaning of “feedback” seemed to be different from that in the MG group, where learners selected and then checked the answer through “feedback.” This point was well taken, especially in light of behavioral differences between the two groups: one simply looks at the answer while the other has to select one. The author is arguing here that the cognitive processes the SG group goes through could be similar to those the MG group does. The SG group could guess the meaning, think about the possible answer, and then confirm it with the answer given in the gloss. In this process, the SG group is getting feedback from the gloss. The author recognizes that this is based on the assumption that the learners in the SG group did guess the meaning before locating the answer and that this assumption needs to be investigated further in future studies.

3. As was mentioned in the “Reading material” section, the reading text with glosses was created using the cloze-making function in Moodle. This was decided in order to conduct the study with the current existing tool without having to write any code. The researcher hoped that this would be an encouragement to those who wish to conduct a similar study without having knowledge of computer programming. However, this has brought a problem of getting access to the feedback page after submitting all the multiple-choice gloss work. An anonymous reviewer pointed out that the timing of getting feedback for the SG and the MG groups should have been controlled. The author acknowledges that this is also a limitation of this study.

4. One could speculate that these six students had already known the words and therefore did not look up the words. However, the pretest results of the students revealed that they hardly knew the words (M=1.83, SD=1.47), and this was quite similar to the results of the students who looked up most of the words (M=1.93, SD=1.94). These six students ignored the glosses for some reasons. It is difficult to tell whether or not this was intentional. The reasons why some learners hardly looked up words remain unknown.

5. The data with all the participants (N = 41) including the six students were also analyzed.
For Pretest, no significant difference was found between the SG (n = 20, M = 1.90, SD = 1.77) and the MG (exactly the same as the one listed in the main analysis): t = .38, DF = 39, p = .70. For Immediate test, the SG had M = 13.90 (SD = 1.17). The difference between the two groups was not significant: t = 1.85, DF = 39, p = .07. However, the effect size represented by Cohen’s d was .59, medium effect. For Delayed test 1, the SG had M = 13.25 (SD = 1.89). The difference between the groups was significant: t = 2.80, DF = 39, p = .008, d = .90, large effect. For Delayed test 2, the SG had M = 11.75 (SD = 1.92), and the significant difference still remained: t = 2.16, DF = 39, p = .037, d = .69 medium to large. The overall results of the analyses corresponded to those in the main analyses except for the immediate test. The data with all the participants did not reveal the significant difference but the data with only those who looked up enough did show the significant difference.

REFERENCES


