Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2023, Volume 27, Issue 1
pp. 1–24

ARTICLE

Multimodal glosses enhance learning of Arabic vocabulary

Juman Al Bukhari, *University of North Georgia*

John A. Dewey, *University of North Georgia*

Abstract

In second language acquisition, a popular method of introducing new vocabulary is by embedding the words in a natural text. Supplementary information (e.g., definitions, illustrations, synonyms, etc.), or glosses, can be included in the margins of the texts to highlight and improve retention of the new words. Previous studies suggest multimodal glosses facilitate learning, although the question of which glosses are most effective remains an active research topic. In the present study, we focused on a sample of university students studying Arabic as a second language. In two experiments, we found that (a) recognition and recall of target vocabulary words were superior when target words were accompanied by a combination of text and picture glosses compared to text-only or picture-only; and (b) including information about the Arabic root word from which the target word was derived in the glosses provided an additional memory benefit. Overall, this study adds further support for glosses as a teaching tool by generalizing to the case of Arabic. Our results also show how glosses that highlight root information can be useful for improving recognition and recall of Arabic vocabulary specifically.

Keywords: *Arabic, SLA, Glosses, Multimodal*

Language(s) Learned in This Study: *Arabic*

APA Citation: Al Bukhari, J., & Dewey, J. A. (2023). Multimodal glosses enhance learning of Arabic vocabulary. *Language Learning & Technology*, *27*(1), 1–24. https://hdl.handle.net/10125/73498

Introduction

In second language acquisition (SLA), a nearly universal challenge for instructors is finding creative ways to help students master new vocabulary. One learning tool that has received a large amount of attention in recent decades is the use of *glosses*, supplementary information such as definitions, illustrations, or synonyms of new vocabulary words, which can be presented in the margins of texts, helping the reader to assimilate the meaning of new words as they read. This technique is partly inspired by theories of knowledge representation imported from cognitive psychology (Paivio, 1990), which propose that comprehension and retention are improved when students encode new information in more than one modality (e.g., verbally and visually, rather than just verbally).

Broadly speaking, the literature suggests that glosses can indeed facilitate learning vocabulary from natural texts (e.g., Chun & Plass, 1996; Gettys et al., 2008; Kost et al., 1999). However, the question of which glosses are the most effective remains an active research topic, and differences across languages, testing environments, and individual learners indicate a continuing need for successful demonstrations in a variety of contexts and populations to maximize the potential for successful application of these findings to real-world classrooms.

In this article, we describe applied research on the impact of glosses in the context of a university classroom where English-speaking students were exposed to new vocabulary through text-based stories in Modern Standard Arabic.[1] Target vocabulary terms were supplemented with different types of gloss information. Besides demonstrating a successful application of glosses in the classroom, we confirm a prediction from

dual coding theories of knowledge representation that multimodal glosses lead to stronger memories than unimodal glosses (text-only or picture-only). We also examined the utility of a novel type of gloss, based on root words, which was specifically geared toward students of Arabic.

Literature Review

Students Can Learn Vocabulary Incidentally through Natural Texts

The procedure by which new vocabulary was presented in this study was inspired by the idea of incidental learning. According to Krashen's (1989) influential Input Hypothesis, SLA is necessarily a scaffolded process in which learners advance by comprehending language that is slightly beyond their current level of competence. This implies that new vocabulary items should be presented as part of a comprehensible input, so that the new material can be related to existing knowledge. For this reason, a popular method of exposing students to new vocabulary is through stories, newspaper articles, and other natural texts. In contrast to presenting students with lists of vocabulary terms to memorize in isolation, natural texts provide context and understanding that may promote incidental (i.e., relatively non-effortful) learning as the learners focus most of their conscious attention on comprehending the broader narrative of the text (Hulstijn & Laufer, 2001; Hunt & Beglar, 2002; Krashen, 1989).

Glosses Supplement Natural Texts

Although natural texts provide contextual clues to the meaning of new words, students may still wish to look up unfamiliar words in a dictionary as they are encountered. Unfortunately, this may disrupt the flow of reading. A less disruptive and time-consuming alternative is for instructors to include definitions and other complementary information (perhaps synonyms or illustrations) in the margins of the text, or with embedded hyperlinks in the case of computer-based presentations. These glosses facilitate uninterrupted reading (Gettys et al., 2008; Laufer & Hill, 2000) while possibly enhancing vocabulary acquisition and reading comprehension (Azari et al., 2012; Choi, 2016; Ko, 2005).

Research on glosses in the context of SLA has often focused on which types of glosses are most optimal for learning. Several studies have reported an advantage for multimodal text+picture glosses compared to text-only or picture-only glosses (e.g., Zarei & Mahmoodzadeh, 2014; Chun & Plass, 1996; Kost et al., 1999; Plass et al., 1998; Yoshii, 2006; Yoshii & Flaitz, 2013). Such results have generally been explained in terms of multimodal theories of knowledge representation. For instance, dual coding theory proposes that memories are represented verbally as well as visually (Paivio, 1990). When a learning event includes both visual and verbal aspects, these features are associated so that each potentially acts as a retrieval cue for the other. For example, a student who struggles to recall a verbal definition may find it easier if they can recall the picture that was shown alongside the word.

Another mechanism which may help explain the beneficial effects of glosses on learning from text is task-induced involvement. According to the Involvement Load Hypothesis, the amount of vocabulary acquired incidentally depends on participants' engagement with the material (Hulstijn & Laufer, 2008). Glosses may cause participants to slow down while reading to think about their relationships to new vocabulary words. This would also be consistent with classic memory research indicating that people are more likely to remember presented words if they process the meaning of those words, rather than just their superficial features (Craik & Tulving, 1975).

How Generalizable are the Effects of Glosses?

Glosses, particularly multimodal glosses, are generally agreed to be a useful educational tool, but it is important to consider how learning is operationalized, as many SLA instructors deploy a combination of recognition, cued recall, and reading comprehension tasks to measure student learning. In addition, glosses could be shown in the learners' first language (L1) or second language (L2).

Indeed, a review of the literature suggests the benefits of glosses may not be equally apparent for all types

of memory tests. For example, in a study of Japanese L2 learners of English, Yoshii (2006) found an advantage for multimodal text+picture glosses over text-only on a cued recall task in which participants were provided the L2 word and had to supply an L1 definition. However, no significant difference was found between gloss types on a recognition task which required selecting the correct definition for an L2 word from four alternatives. On the other hand, in a study of English L2 learners of Spanish, Yanguas (2009) found an advantage for text+picture glosses on tests of reading comprehension but not on a production test in which participants were given a list of English words and had to write their Spanish equivalents. These inconsistencies and caveats highlight the importance of using multiple measures of learning to provide convergent evidence of the effectiveness of glosses. They also highlight the importance of testing glosses in a variety of contexts to better understand the limits of the effects.

Individual differences between learners may also be an important parameter to consider regarding the effectiveness of glosses. As previously noted, glosses can provide information to the learner in either their L1 or L2. An L2 gloss bypasses the translation step, so learners directly associate a new vocabulary word to its concept. This contrasts with the strategy of learning a new L2 word by associating it to an L1 word, and only then activating the concept. Theoretical accounts of knowledge representations in the minds of language learners suggest that advanced learners have a better ability to directly link L2 words to concepts than novices do (Kroll & Stewart, 1994; Kroll & Sunderman, 2003; Potter et al., 1984). This predicts that L2 root glosses may be more effective for advanced students, while L1 glosses might be better for beginning students. Indeed, previous research suggests that, at least in some contexts, L2 glosses may be more effective than L1 glosses for advanced learners, while the opposite is true for beginners (Ko, 2005; Miyasako, 2002).

Another individual difference that could potentially moderate the effects of glosses is working memory capacity (i.e., the amount of information an individual can hold and manipulate in conscious memory at one time). People with greater capacity tend to have an advantage in many settings, including assessments of reading comprehension and standardized academic tests (e.g., Ackerman et al., 2002; Blankenship et al., 2015; Engle & Kane, 2004; Hannon & Daneman, 2001). One could imagine that people with a lower working memory capacity might be more easily overwhelmed or distracted if too much information were presented in the glosses.

Another variable that could potentially interact with gloss effects would be the testing environment itself. A popular topic in education and language research in recent decades has been computer-assisted language learning (CALL). Increasing access to technology in the classroom gives instructors many more options for supplementing text with other resources, such as hyperlinks or videos. For example, a meta-analysis by Taylor (2009) found larger gloss effects for CALL studies than for non-CALL studies. In the present study, we did not compare CALL and non-CALL methods of presenting glosses, but we did consider the effects of being tested on a computer versus being tested using a traditional pen-and-paper test. Different testing modes could potentially influence the results for at least two reasons. First, the degree of overlap between conditions during study and test can influence how easily information is retrieved during the test (e.g., Godden & Baddeley, 1975). In the present study, all our participants were first exposed to the texts containing the target vocabulary on computers, so one might predict better test scores if the recognition or recall tests were also administered on computers. Second, the amount of time that students spent producing their responses could also influence their accuracy (the speed-accuracy tradeoff). For example, if writing one's answers on a pen-and-paper test took a student more time compared to typing the same answers using a computer keyboard, that extra time spent reflecting on the answer might improve their accuracy. Further important considerations relating to the effectiveness of glosses may be language specific. In the present study, we focused on English-speaking L2 learners of Arabic. Therefore, it is useful to briefly consider some properties of the Arabic language that may present challenges for students. These considerations influenced the manipulations used in Experiment 2.

Challenges for L2 Learners of Arabic

There are several obstacles that may hinder readers of Arabic texts, including diglossia (i.e., two variations

of the same language used in the same speech community, but with distinct functions; Ferguson, 1959), homographic words, morphological and phonological complexity, and short vowelization of words as a function of their grammatical functions in sentences (Abu-Rabia, 2019). For example, many words look the same unless short vowels are placed either on top of or underneath the letter, which helps the reader to disambiguate the words. Advanced learners can read texts even without the presence of marked short vowels (deep orthography), using context to infer the short vowels and thus the meaning of the words. However, this vowel length ambiguity can pose a significant challenge for less advanced learners.

Arabic uses both derivational and inflectional morphology. The former is based on phonological patterns constructed on roots that are consonant patterns. For instance, *kaatib* ('writer') is derived from the root *k-t-b* ('to write'). Similarly, other derivations from the same root share a semantic relationship to writing (see Table 1). The inflectional morphology involves attaching prefixes and suffixes to nouns and verbs. The Arabic verb system inflects for phi-features (gender, person, and number) and tense (or time), while the Arabic noun system inflects for gender and number.

To create new words and change meaning from a root like *k-t-b*, different vowels are inserted as in *kitaab* ('book'), lengthened as in *kaatib* ('writer'), or consonant sounds are inserted at the beginning of the word as in *maktab* ('office'). The verb *takaataba* ('write to each other') is also an example of an inflection for phi-features.

Table 1

Example of an Arabic Root (k-t-b), Related Words, and English Equivalents

Arabic	English
k-t-b	to write
kitaab	book
kaatib	writer
kitaaba	writing
maktab	office
maktabe	library
takaataba	write to each other
maktuub	written

When learners of Arabic come across a new word such as *kaatib* ('writer'), they may infer its relation to *kitaaba* ('writing') if they know the root *k-t-b*, or the singular stem word *kitaab* ('book'). For this reason, Arabic language educators commonly remind students of the derivational origins of new words to help students categorize and remember vocabulary. This practice aligns well with the theoretical perspective that language and concept learning influence one another and naturally develop side-by-side in children (e.g., Borovsky & Elman, 2006).

To summarize, reading Arabic without marking short vowels is cognitively demanding, especially for beginning students. However, it is possible that glosses could help facilitate this process. To our knowledge there have been no empirical studies investigating the effects of multimodal glosses on vocabulary acquisition for Arabic specifically. Moreover, we were interested in the effects of supplementing Arabic texts with glosses indicating the root words for new vocabulary terms. We reasoned that for L2 Arabic instruction specifically, this information would benefit learning by drawing students' attention to the shared connections of the semantic unit (a pedagogical goal unto itself), while also benefiting encoding and retrieval of the new vocabulary words by connecting new knowledge to existing knowledge. On the other

hand, providing too much information simultaneously can split attention and potentially be overwhelming to learners (see Holsanova et al., 2009 for general information about the split-attention effect; see Taylor, 2009, 2010 for discussion of how glosses can be distracting), so the outcome of this intervention was not obvious in advance.

Purpose

In the present study, we examined the benefits of multimodal glosses in the context of a college classroom using text-based stories to promote incidental learning of Arabic vocabulary. Our purposes were: (a) To conceptually replicate and generalize previous studies that found an advantage for multimodal glosses (text+picture vs. text-only or picture-only) while accounting for potential individual differences related to students' working memory capacity; and (b) To test the hypothesis that, for students of Arabic specifically, providing the L2 root (such as 'k-t-b') as a gloss would further boost learning of the new vocabulary.

Research Questions

1. Do multimodal glosses lead to superior recall and recognition of Arabic vocabulary, compared to unimodal glosses?
2. Does providing L2 root information in glosses provide an additional benefit for students of Arabic?

Experiment 1

In our first experiment, we attempted to conceptually replicate previous studies that found improved recognition and/or recall of vocabulary presented alongside multimodal glosses, compared to similarly difficult vocabulary presented alongside unimodal glosses. This was a necessary first step before we could address our more novel research question in the second experiment. To our knowledge, there are no published studies demonstrating that multimodal glosses improve recognition and recall of Arabic vocabulary. With Experiment 1, we set out to establish that our procedure was adequate to replicate known gloss effects, while also gathering information related to individual differences in working memory capacity that we suspected might interact with the effects of the glosses.

Based on our literature review, we hypothesized that students who were exposed to new vocabulary via text-based stories would be more likely to recognize and recall new words that were supplemented by multimodal glosses (text+picture) in the margins, compared to words with unimodal glosses (text or picture only). To potentially improve the generalizability of our results, we also manipulated the way students were tested (computerized test vs. pen-and-paper), and the retention interval between study and testing events, although we had no specific hypotheses related to these manipulations other than the obvious prediction that participants would perform better at the shorter retention interval. Finally, we also measured participants' reading span, a measure of verbal working memory capacity (Baddeley, 2010; Daneman & Carpenter, 1980), which we considered as a potential moderating factor with respect to the effects of the glosses.

Method

Participants

We recruited a convenience sample consisting of 41 undergraduate L2 learners of Arabic at a medium-sized public university in the southern United States. Participants were selected from four different course sections representing different levels of proficiency (novice, intermediate-low, intermediate-high, and advanced). Participants' level depended on the number of hours covered in learning Arabic as an L2 language at a Summer Language Institute held at the university. The correspondence between hours of training and proficiency was based on the ACTFL Oral Proficiency Interview (OPI) ratings, which represent levels of expected performance for language learners who complete full-time intensive and/or immersion, proficiency-based language training. For example, 480 hours of training corresponds to an average expected performance of "intermediate-low", while 720 hours of training corresponds to an average expected

performance of "intermediate-high". The first author attempted to recruit average students to represent each of the different levels, as even students within the same class/level may display large differences in proficiency. Overall, 10 were drawn from the novice level, 5 were intermediate-low, 19 were intermediate-high, and 7 were advanced. The instructor was held constant.

Design

We used a 2 x 2 x 3 within-subjects factorial design. Because of our limited sample size, a within-subjects design was used to control for individual differences. The manipulated variables were: (a) Testing mode (computer vs. pen-and-paper); (b) Delay (immediate or two weeks) between the study and test phases; and (c) Gloss (text, picture, or text+picture) condition. The order of the testing modes was counterbalanced across participants, such that half of the participants were initially tested on computers and later tested using pen-and-paper tests, and the other half were tested in the reverse order. Different texts and target vocabulary words were used for each of the posttests.

Notably, our design did not include a true control condition with no glosses. This was partly because the instructor wanted students to have a fair chance to learn all the target words, and with no explicit L1 definitions or glosses, the meanings of some words would be difficult to decipher from context alone.

The dependent variable was the percentage of correct responses on subsequent tests of recognition and production. We also measured participants' reading span scores and proficiency levels (course section).

Materials

Background Questionnaire. Participants were asked to complete a background questionnaire that included questions about the number of years spent studying Arabic (recently or as a child), self-ratings of ability to speak, read, and understand spoken Arabic, and whether they had second language experience other than Arabic (see Appendix A).

Pretest. A pretest was used to assess participants' knowledge of target vocabulary items prior to the interventions. The pretest consisted of a written list of 28 words (21 total target words with 7 from each condition, and 7 non-target words; see Appendix B for sample items on the pretest). The instructions for the pretest asked participants to provide the meaning of any words they already knew, or to write 'NS' (not sure) if they were not sure of the meaning. Because our study was focused on how students learn and retain new vocabulary, we planned to exclude any participants from later analyses if they already knew 20% or more of the target words at pretest. However, no participants scored at or above 20% on the pretest.

Texts. The stimuli were passages from an online English-language newspaper that had previously been translated into Arabic. There were eight total passages, with two passages used for each level of student (one used for the computerized tests, the other for the pen-and-paper test). Each passage included 150–160 words of which 21 words (the target words) were glossed, with seven glossed words per condition (i.e., seven words with text glosses, seven with picture glosses, and seven with a text+picture glosses). See Appendix C for examples of sentences from these passages which show how glosses were presented alongside the text, and Appendix D for a table showing examples of target words used in each condition. Participants accessed the texts through the university's course management software (D2L) using iPad tablets provided by the language lab.

Posttests. To measure participants' retention of target vocabulary words (the words with glosses), two types of posttest were used, a recognition test and a production test. In the recognition test, participants were asked to pick out the English equivalent of an Arabic word from four alternatives (see Appendix E for examples). In the production task, participants were asked to produce (write) the Arabic equivalent of an English phrase (see Appendix F for examples). For pedagogical reasons, a reading comprehension test was also administered which asked general questions about the passages that participants had read. However, the reading comprehension tests did not play a role in our present hypothesis testing because they measured participants' understanding of the story as a whole, rather than specific words linked to the different gloss conditions. Therefore, the reading comprehension tests will not be discussed further in this paper.

Working Memory Test. Each participant in Experiment 1 completed a computerized implementation of the reading span task to measure their working memory capacity (using the Psychology Experiment Building Language, see Mueller & Piper, 2014). Like the original offline version (Daneman & Carpenter, 1980), this implementation of the reading span task contains a processing component in which the participants judge the semantic correctness of a sentence, and a storage component where the participants must memorize a list of letters in the correct order for later recall. On each trial of the reading span test, participants were shown a sequence of letters to hold in memory, such as [*A*, *T*, *C*, *O*], then shown a sentence, which they had to judge as sensible (e.g., "They like to run in the park.") or nonsense (e.g., "I like to run in the sky."). Following the semantic judgment, participants would try to report the letter sequence. Participants were tested using letter sequences up to a maximum length of seven. Participants' reading spans were determined by fitting a psychometric function to their data and calculating the list length at which 80% accuracy was expected. The analysis was performed using the 'psyphy' package of functions in Version 3.3.1 of the R statistical computing language (R Core Team, 2016).

Procedure

Data collection took place during regularly scheduled class meetings in the university's language lab, which provided a quiet test environment. Each participant's data was collected over the span of five class meetings.

During the first meeting, participants completed the background questionnaire and the pretest to assess prior knowledge of target vocabulary words, studied the text containing the target words, and took the first, immediate set of posttests (recognition, production, and reading comprehension) that were either administered by computer (half the participants) or pen-and-paper (the other half of participants). Participants were given approximately 15–20 minutes to study the passage before beginning the posttests.

Participants were asked not to search for any information related to the tested vocabulary items during the delay period. Two weeks after the first posttests, participants were tested again using the same procedure as the immediate posttest. Participants completed the delayed posttests in the same test modality (computer or pen-and-paper) as the immediate tests. The delayed tests were the same as the immediate tests, except the order of the questions and the response options were scrambled.

At the third meeting, participants had their reading span scores measured. On each trial of the reading span task, a sequence of letters appeared at the center of the screen. Each letter was displayed for 500 ms. The number of letters varied between two and seven. At the end of the letter sequence, a blank screen was shown for 500 ms, then a sentence was shown in the center of the screen for a maximum of 10 seconds or until the participant provided a response. The average length of the sentences was 5.4 words. Participants were asked to read the sentence and indicate if it made sense or not. The participant hit '1' if it made sense and '0' if it did not make sense. After another blank screen of 500 ms, the participants were asked to enter the initial list of letters they saw in the same order they appeared. There were 46 total trials in the reading span task.

In the fourth and fifth class meetings, students were given a new passage and repeated the first two steps, but in the other test modality. Those who were initially tested with pen-and-paper were tested on the new passage with computerized tests, and vice versa.

Results

Before the main analyses, we checked the correlations between reading span scores and scores on the pretest, recognition, and production tasks. Most of the correlations were positive but small in magnitude, and none were statistically significant. Therefore, we do not consider reading span any further in our analysis of the results. We combined the data from students of different class levels (novice, intermediate-low, etc.).

The recognition and production tasks were analyzed with a pair of 2 (Testing Mode: computer vs. pen-and-paper) x 2 (Delay: immediate vs. delayed) x 3 (Gloss: text vs. picture vs. text+picture) repeated-measures analyses of variance (ANOVA). When post hoc *t*-tests were used, we used Bonferroni *p*-value adjustment to correct for multiple comparisons.

For all tests, we used an alpha of .05 as our threshold for statistical significance, and partial eta squared (η^2_p) as our measure of effect size.

Pretest

On average, participants were able to provide the meaning for 2.32 out of 25 target words (about 9%) prior to reading, with a standard deviation of 1.8. No participants were excluded from further analysis based on the results of the pretest.

Recognition Task

Descriptive statistics for the recognition task are summarized in Table 2, and visually summarized in Figure 1.

Table 2

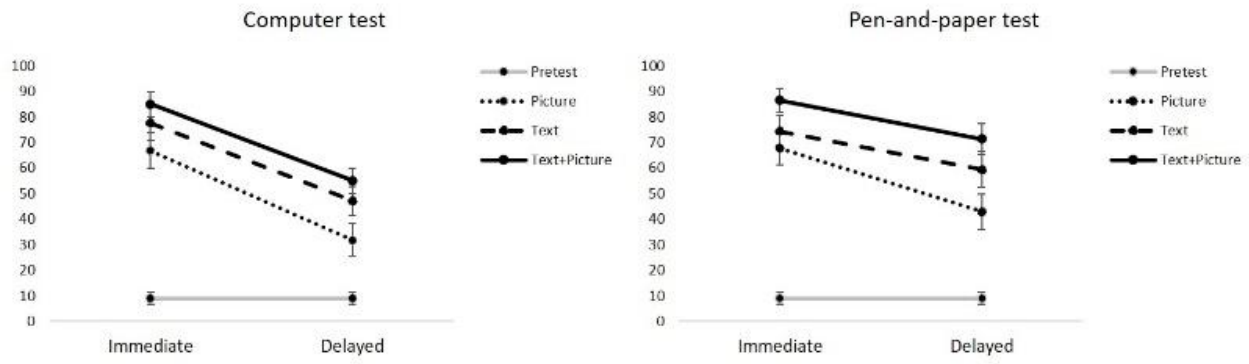
Descriptive Statistics for the Recognition Task, Experiment 1 (N = 41)

Test mode	Delay	Gloss	*M* (%)	*SD*
Computer	Immediate	Text	77.68	22.19
		Picture	66.83	22.88
		Text+Picture	85.02	16.46
	Delayed (2 weeks)	Text	47.17	17.95
		Picture	31.83	21.32
		Text+Picture	55.07	16.25
Pen-and-Paper	Immediate	Text	74.30	21.40
		Picture	67.73	21.30
		Text+Picture	86.60	14.80
	Delayed (2 weeks)	Text	59.46	22.98
		Picture	43.01	22.47
		Text+Picture	71.40	20.03

Note. Means indicate % correct responses.

Figure 1

Means and 95% Confidence Intervals for the Recognition Task, Experiment 1



The main effect of Testing Mode was significant, $F(1, 40) = 6.10$, $p = .02$, $\eta^2_p = .13$, indicating that participants performed better on the pen-and-paper tests ($M = 67.08$, $SE = 2.29$) compared to the computer tests ($M = 60.60$, $SE = 2.29$).

There was also a significant main effect of Delay, $F(1, 40) = 161.67$, $p < .001$, $\eta^2_p = .80$, indicating that recognition scores were higher on the immediate posttest ($M = 76.36$, $SE = 2.12$) compared to the delayed posttest ($M = 51.32$, $SE = 2.12$).

There was also a significant main effect of Gloss, $F(2, 80) = 121.12$, $p < .001$, $\eta^2_p = .75$. As shown in Figure 1, recognition scores were highest for text+picture glosses ($M = 74.42$, $SE = 2.05$), followed by text glosses ($M = 64.65$, $SE = 2.05$), then picture glosses ($M = 52.35$, $SE = 2.05$). Post hoc *t*-tests showed that all three levels were significantly different from each other with $p < .05$.

The two-way interaction between Testing Mode and Delay was also significant, $F(1, 39) = 9.26$, $p = .004$, $\eta^2_p = .19$. The mean difference between the immediate and delayed posttest was larger for the computer tests (31.82) than for the pen-and-paper tests (18.26). This indicates more forgetting over time for the computer tests.

The two-way interaction between Delay and Gloss was also significant, $F(2, 80) = 3.53$, $p = .034$, $\eta^2_p = .08$. The difference between the immediate and delayed tests was larger for the picture glosses (29.86) than for text (22.67) or text+picture (22.58), indicating more forgetting when picture glosses were used.

Production Task

Descriptive statistics for the production task are summarized in Table 3, and visually summarized in Figure 2.

Table 3

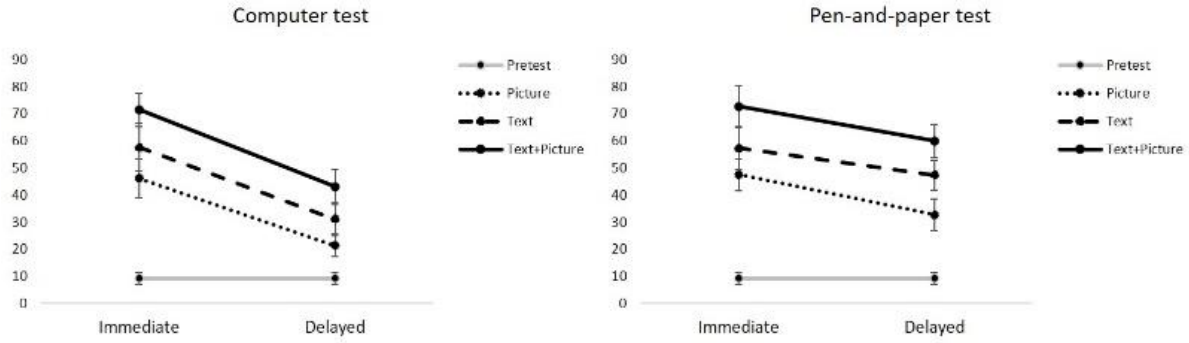
Descriptive Statistics for the Production Task, Experiment 1 (N = 41)

Test Mode	Delay	Gloss	*M*(%)	*SD*
Computer	Immediate	Text	57.66	28.29
		Picture	46.20	23.83
		Text+Picture	71.46	20.43
	Delayed (2 weeks)	Text	31.15	19.74
		Picture	21.32	13.44
		Text+Picture	43.07	21.36
Pen-and-Paper	Immediate	Text	57.29	25.28
		Picture	47.53	19.58
		Text+Picture	72.71	24.44
	Delayed (2 weeks)	Text	47.35	17.97
		Picture	32.69	19.07
		Text+Picture	60.00	20.43

Note. Means indicate % correct responses.

Figure 2

Means and 95% Confidence Intervals for the Production Task, Experiment 1



There was a significant main effect of Testing Mode, $F(1, 40) = 8.45$, $p = .006$, $\eta^2_p = .17$, indicating that participants performed better on the pen-and-paper tests ($M = 52.93$, $SE = 2.35$) compared to the computer tests ($M = 45.14$, $SE = 2.35$).

The main effect of Delay was also significant, $F(1, 40) = 67.37$, $p < .001$, $\eta^2_p = .63$, indicating that participants performed better on the immediate test ($M = 58.81$, $SE = 2.27$) compared to the delayed test ($M = 39.26$, $SE = 2.27$).

There was also a significant main effect of Gloss, $F(2, 80) = 127.54$, $p < .001$, $\eta^2_p = .76$. As can be seen in Figure 2, production scores were highest for text+picture glosses ($M = 61.81$, $SE = 2.13$), followed by text glosses ($M = 48.36$, $SE = 2.13$), then picture glosses ($M = 36.93$, $SE = 2.13$). Post hoc *t*-tests showed that all three conditions were significantly different from the others with $p < .05$.

Finally, the two-way interaction between Delay and Test Mode was significant, $F(1, 40) = 9.62$, $p = .004$, $\eta^2_p = .19$. The mean difference between the immediate and delayed posttest was larger for the computer tests (26.59) than for the pen-and-paper tests (12.50). This indicates more forgetting over time for the computer tests.

Discussion

The main purpose of Experiment 1 was to conceptually replicate previous studies that showed a memory advantage in SLA for vocabulary words enhanced by multimodal glosses to ensure that the effect generalizes to learning Arabic. Consistent with previous literature, our participants performed the best in the text+picture gloss condition compared to conditions with text or picture glosses only. This multimodal advantage was apparent for recognition and production tasks, on both immediate and two-week delayed tests, suggesting a robust effect.

We were also curious if reading span scores were associated with vocabulary acquisition on our task. If so, it might be necessary to include reading span as a covariate in our analysis to control for individual differences. However, reading span scores did not significantly correlate with recognition or production scores.

An unexpected finding from Experiment 1 was that participants performed better on pen-and-paper tests compared to the computerized tests after a two-week delay. The target vocabulary items were counterbalanced across Testing Mode, so the pen-and-paper and computer tests should not have differed in terms of their difficulty. Additionally, all participants completed a computerized test as well as a pen-and-paper test, so the difference across test modes cannot be attributed to individual differences between students. Although it is beyond the scope of the present study, a speculative explanation is that a pen-and-paper test may take longer to complete than a computer test, assuming that typing is faster for many people

than writing longhand. If so, perhaps this slower pace is beneficial for entering that same material into long-term storage. Another explanation could be that the students in our sample previously had more practice writing the Arabic alphabet longhand compared to typing Arabic characters on the computer. This would not be unusual since Arabic does not use the same Latin alphabet as English, and many students do not have extensive experience using Arabic keyboards.

Experiment 2

Having replicated the memory advantage for vocabulary paired with multimodal glosses, our next goal was to boost performance even further. Specifically, we hypothesized that L2 root information, which indicates the semantic origin of Arabic vocabulary words, would provide an additional benefit when paired with any of the gloss conditions tested in Experiment 1.

Method

Participants

We sampled 35 undergraduate students studying Arabic as an L2 language at a medium-sized public university in the southern United States, with 23 participants from the intermediate-high level class and 12 from the advanced class. As in Experiment 1, the level of the students was based on the number of hours they previously completed at a Summer Language Institute held at our university. We excluded novice students because, at the time of testing, they had not acquired enough vocabulary words of the same root, which made it difficult to create appropriate test stimuli for those students. The instructor was held constant.

Design

We used a 3 x 2 x 2 within-subjects factorial design. The independent variables were: (a) Base Gloss (text, picture, or text+picture); (b) Delay (immediate or two weeks) between the study and test phases; and (c) Root (a root word was included with the base gloss, or was not). As in the first experiment, the main dependent variables were the percentage of correct responses on subsequent tests of recognition and production.

Materials and Procedure

The materials and procedure were similar to those for the first experiment, except that we did not test participants' reading span, and we removed the Test Mode variable from Experiment 1. All participants were tested using a traditional pen-and-paper format.

As in Experiment 1, participants completed a pretest to measure their prior knowledge of target vocabulary items. We intended to exclude any participant whose score surpassed 20%. However, no participants had to be excluded.

Data collection took place during two regularly scheduled class meetings per group (intermediate and advanced) in the university's language lab. At the first meeting, participants completed a background questionnaire, a pretest to assess prior knowledge of target vocabulary words, studied the text containing the target words, and took the immediate recognition and production posttests. As in Experiment 1, there were seven glossed target words per condition. Two passages were used, one for the intermediate students and one for the advanced students (See Appendix G for examples). At the second meeting, participants completed the delayed posttests. As in Experiment 1, the delayed tests were identical to the immediate tests except the question order and response options were scrambled.

Results

The recognition and production tasks were analyzed with a pair of 3 (Base Gloss: text vs. picture vs. text+picture) x 2 (Delay: immediate vs. delayed) x 2 (Root: root vs. no root) repeated-measures analyses of variance (ANOVA). All results include both the intermediate-high and advanced students. When post hoc *t*-tests were used, we used Bonferroni *p*-value adjustment to correct for multiple comparisons.

Pretest

On average, participants were able to provide the meaning for 2.6 out of 25 target words (about 10%) prior to reading, with a standard deviation of 1.54. No participants were excluded from further analysis based on the results of the pretest.

Recognition Task

Descriptive statistics for the recognition task are summarized in Table 4, and visually summarized in Figure 3.

Table 4

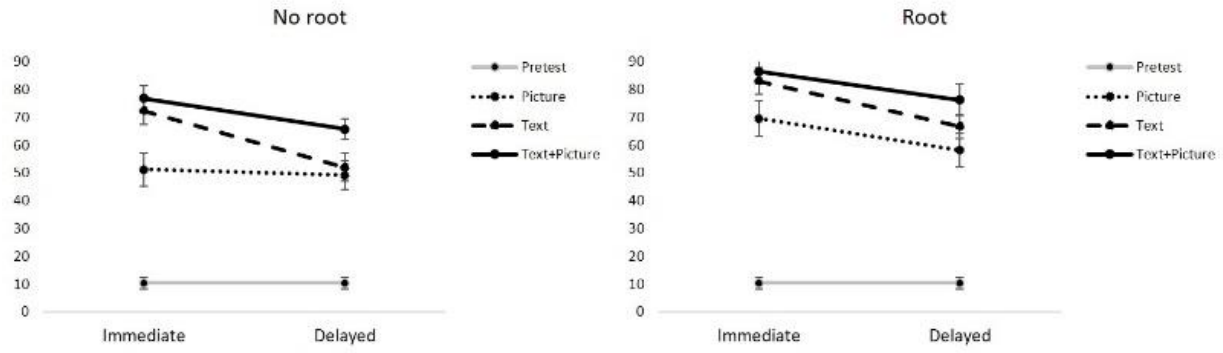
Descriptive Statistics for the Recognition Task, Experiment 2 (N = 35)

Base Gloss	Root	Delay	*M*(%)	*SD*
Text	No root	Immediate	72.43	15.27
		Delayed (2 weeks)	52.00	15.06
	Root	Immediate	83.09	14.08
		Delayed (2 weeks)	66.71	13.05
Picture	No Root	Immediate	51.17	17.79
		Delayed (2 weeks)	49.17	15.82
	Root	Immediate	69.63	18.93
		Delayed (2 weeks)	58.14	18.61
Text+Picture	No Root	Immediate	76.80	13.49
		Delayed (2 weeks)	65.74	10.94
	Root	Immediate	86.40	12.91
		Delayed (2 weeks)	76.29	16.90

Note. Means indicate % correct responses.

Figure 3

Means and 95% Confidence Intervals for the Recognition Task, Experiment 2



The main effect of Delay was significant, $F(1, 34) = 37.21$, $p < .001$, $\eta^2_p = .52$, indicating that recognition accuracy was higher on the immediate test ($M = 73.25$, $SE = 1.72$) compared to the delayed test ($M = 61.34$,

SE = 1.72).

The main effect of Base Gloss was also significant, $F(2, 68) = 64.31$, $p < .001$, $\eta^2_p = .65$. Post hoc *t*-tests showed that all three Base Gloss types were significantly different from each other (all with $p < .001$), with the highest recognition accuracy for text+picture (*M* = 76.31, *SE* = 1.73), followed by text (*M* = 68.56, *SE* = 1.73), then picture (*M* = 57.03, *SE* = 1.73).

The main effect of Root was also significant, $F(1, 34) = 133.23$, $p < .001$, $\eta^2_p = .80$, indicating higher recognition scores for the root condition (*M* = 73.38, *SE* = 1.51) compared to the no root condition (*M* = 61.22, *SE* = 1.51).

The interaction between Base Gloss and Delay was also significant, $F(2, 68) = 5.10$, $p < .01$, $\eta^2_p = .13$. The difference between the immediate and delayed tests was largest for text (18.4), followed by text+picture (10.59), then picture (6.74).

Production Task

Descriptive statistics for the production task are summarized in Table 5, and visually summarized in Figure 4.

Table 5

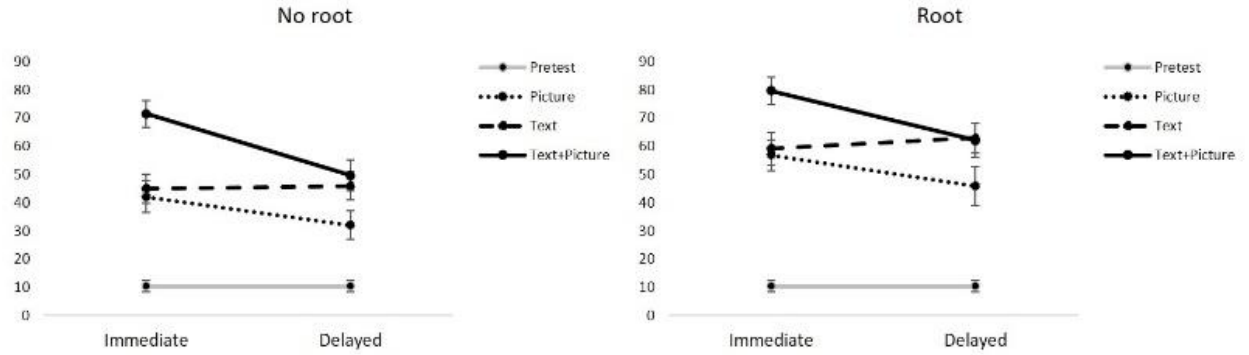
Descriptive Statistics for the Production Task, Experiment 2 (N = 35)

Base Gloss	Root	Delay	*M*(%)	*SD*
Text	No root	Immediate	44.97	15.46
		Delayed (2 weeks)	45.89	15.13
	Root	Immediate	59.17	17.20
		Delayed (2 weeks)	62.97	15.98
Picture	No Root	Immediate	42.06	16.89
		Delayed (2 weeks)	31.97	15.26
	Root	Immediate	56.80	16.20
		Delayed (2 weeks)	45.86	20.70
Text+Picture	No Root	Immediate	71.46	13.93
		Delayed (2 weeks)	49.66	16.59
	Root	Immediate	79.66	14.60
		Delayed (2 weeks)	62.06	18.19

Note. Means indicate % correct responses.

Figure 4

Means and 95% Confidence Intervals for the Production Task, Experiment 2



The main effect of Delay was significant, $F(1, 34) = 16.53$, $p < .001$, $\eta^2_p = .33$. This indicates that production accuracy was higher on the immediate test ($M = 59.02$, $SE = 1.91$) compared to the delayed test ($M = 49.73$, $SE = 1.91$).

The main effect of Base Gloss was also significant, $F(2, 68) = 52.27$, $p < .001$, $\eta^2_p = .61$. Post hoc *t*-tests showed that all three Base Gloss types were significantly different from each other (all with $p < .001$), with the highest production accuracy for text+picture ($M = 65.71$, $SE = 1.96$), followed by text ($M = 53.25$, $SE = 1.96$), then picture ($M = 44.17$, $SE = 1.96$).

The main effect of Root was also significant, $F(1, 34) = 154.95$, $p < .001$, $\eta^2_p = .82$, indicating higher production accuracy for root ($M = 61.09$, $SE = 1.62$) than for no root ($M = 47.67$, $SE = 1.62$).

Finally, the interaction between Base Gloss and Delay was also significant, $F(2, 68) = 19.94$, $p < .001$, $\eta^2_p = .37$. The difference between the immediate and delayed tests was largest for text+picture (19.7), followed by picture (10.52), then text (-2.36).

Discussion

The purpose of Experiment 2 was to explore whether the memory advantage for Arabic vocabulary words enhanced by multimodal glosses could be further boosted with root word information. Our results indicate that root words did indeed improve students' learning of the target vocabulary. This effect was apparent for both the recognition and production tasks, and on both immediate and two-week delayed tests.

We propose that root word glosses could enhance learning for multiple reasons. As with the other types of glosses we tested, root words can help learners by providing additional context to process the meaning of new vocabulary words. If students pause to think about the material in the glosses, this extra engagement with the material should promote deeper processing, which can be advantageous for learning. During the recognition and production tests, associations between glosses and new vocabulary words may have provided students with additional retrieval pathways to access information related to the newly learned words. Beyond that, L2 roots may further improve learning by helping learners understand the derivation of new vocabulary words and their semantic relatedness. For instance, the root *katab* ('to write') organizes the words *maktab* ('office'), *maktabe* ('library'), and *kitaab* ('book') into one category. This improves the organization of the to-be-memorized material, which tends to benefit learning.

General Discussion

Two experiments were used to investigate the effects of supplementary glosses on second language learners' retention of new Arabic vocabulary presented through natural texts. Experiment 1 conceptually replicated previous studies which have shown an advantage of multimodal (text+picture) glosses over

unimodal (text or picture only) glosses. This was important to demonstrate, as most similar studies have concentrated on Latin orthographies, and as previously noted in the literature review, Arabic has a complex morphology that differs from other languages. Experiment 2 demonstrated an additional benefit of including L2 root word glosses with the text.

Beginning with Experiment 1, the superiority of multimodal glosses was not particularly surprising to us, but our results support the generalizability of similar past studies to the present context (college-level Arabic instruction). We further supported the generalizability of this finding by showing that it occurred for pen-and-paper tests as well as computerized tests, for recognition tests as well as production (cued recall) tests, and for both immediate and two-week delayed tests. Like other investigators, we understand this result in terms of multimodal theories of knowledge representation, of which Paivio's (1990) dual coding theory is a well-known early example. These theories assume that memories are represented in multiple codes (visual, verbal, and perhaps other formats as well) which can prime each other through spreading activation during memory retrieval. For example, if a target vocabulary word was associated with a picture as well as a text-based definition, either of those two representations could provide a point of entry for recalling the target word during a later test. Thus, multimodal glosses improve memory by adding redundancy and richness during encoding, and another possible retrieval cue during recall. Another perspective through which our results can be understood is the idea of involvement load. Active engagement with a text (i.e., pausing to think about or elaborate on concepts from the text during reading activities) is better for vocabulary acquisition compared to a more passive reading activity (Hulstijn & Laufer, 2008). In other words, the glosses may have improved memory because they caused participants to engage with the meaning of the new vocabulary words more actively.

Incidentally, we also found that, on average, students performed better on pen-and-paper tests compared to computerized tests. We did not anticipate this result, but it may provide an interesting data point for educators with an interest in the advantages and disadvantages of computerized test environments, some of which might be discipline-specific. For example, we suggested in the Discussion section of Experiment 1 that, for Arabic and other languages that do not use the Latin alphabet, some students might express themselves more fluently (and thus test better) in the traditional pen-and-paper mode. This is not to deny the advantages of technology use in SLA. Applied correctly, technology in the classroom promotes autonomy and student interest in language learning (Beatty, 2005).

Our second experiment tested a different type of gloss which has, to our knowledge, never been reported on prior to this study. Because semantically related words tend to share a common root in Arabic, we hypothesized that glosses which drew attention to the root word would aid memory by relating the new material to existing knowledge. This hypothesis was supported, as the additional root information increased test scores even beyond the advantages observed for multimodal glosses in Experiment 1. Again, this result makes sense in relation to well-accepted theories of memory and knowledge representation (i.e., depth of processing, multimodal knowledge representation, spreading activation theories of memory retrieval). Presenting too much information simultaneously can sometimes overwhelm learners (e.g., Taylor, 2009, 2010), so there may be a limit on how much one can add to glosses before they become a distraction to learning, rather than a complement. However, we did not appear to hit that limit in the present study. A possible caveat to the results of Experiment 2 is that some roots have multiple meanings. In the present study the target vocabulary words were selected to avoid any such issues, and we presume that if ambiguous root information were provided, this would be less helpful for learners.

A potential future direction for this research would be to compare alternative methods of linking the new vocabulary to its semantic category. In Experiment 2 we used roots, but perhaps there could be a similar benefit to glosses showing related stem words. For example, if the target vocabulary word was *kaatib* ("writer") then the gloss could be *kitaab* ("book"). This would seem to achieve a similar aim as glossing the root word *k_t_b*, but the effects could be subtly different if, for example, the connection between a target word and its stem is stronger or more direct than the connection between a target word and its three-consonant root.

In this study we focused on L2 glosses. As noted in the literature review, presenting glosses in L1 or L2 can have implications for student learning as past studies found that L2 glosses were more beneficial for advanced learners (Ko, 2005; Miyasako, 2002). Although we tested students from a variety of proficiency levels, we did not have a sufficiently large sample to meaningfully evaluate whether student proficiency moderated the effect size of the L2 glosses with root words. This could be a potential avenue for future research.

An important limitation of the present research is that we used a convenience sample rather than a random sample. This necessarily limits the generalizability of our results to similar populations and contexts (specifically, teaching Arabic as a second language to university-aged students). At the same time, another important aspect of generalizability is ecological validity, or the realism of the testing situation. In this respect, our study had good validity, since we studied students' performance in a real-world classroom setting.

Our study used a within-group, repeated measures design. A potential downside to this design is that demand characteristics can potentially influence the results. Specifically, when participants are exposed to every level of an independent variable (i.e., the different gloss conditions), there is a risk they may guess the hypothesis of the study and alter their natural behavior as a result. Researchers attempting similar interventions in the future may wish to consider an independent-groups design with random assignment to avoid this potential problem.

Another important limitation is that our study used only a small sample of vocabulary words, and we did not control for features such as the length of the words or pronounceability. This could potentially have confounded our results or obscured differences between the gloss conditions. However, the fact that we found large and consistent differences between the gloss conditions across two experiments, despite modest sample sizes, suggests that multimodal glosses may have practical value as a classroom intervention to help students retain new vocabulary. Finally, neither Experiment 1 nor Experiment 2 included a true control condition without any glosses. As previously noted, this was partly because we did not wish to prevent students from learning certain target words. Nonetheless, the lack of a no-gloss baseline condition means we can only address the relative benefits of different glosses. A no-gloss baseline would be a desirable feature in similar future studies.

We conclude with two main findings. First, this study adds further empirical support regarding the effectiveness of multimodal glosses for promoting incidental vocabulary learning from text. Second, glosses that help students associate Arabic vocabulary words with the roots from which they are derived improve recognition and recall of new vocabulary words beyond what is achieved with definitions and visual illustrations.

Note

1. We adopt the communicative approach in teaching Arabic, where students learn Levantine Arabic for speaking, and Modern Standard Arabic for reading and writing.

References

Abu-Rabia, S. (2019). The role of short vowels in reading Arabic: A critical literature review. *Journal of Psycholinguistic Research, 48,* 785–795. https://doi.org/10.1007/s10936-019-09631-4

Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*(4), 567–589. https://doi.org/10.1037/0096-3445.131.4.567

Azari, F., Abdullah, F. S., Heng, C. S., & Hoon, T. B. (2012). *Effects of glosses on vocabulary gain and retention among tertiary level EFL learners* (ED533228). ERIC. https://files.eric.ed.gov/fulltext/ED533228.pdf

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

Beatty, K. (2005). *Teaching and researching: Computer-assisted language learning* (2nd ed). Routledge.

Blankenship, T. L., O'Neill, M., Ross, A., & Bell, M. A. (2015). Working memory and recollection contribute to academic achievement. *Learning and Individual Differences*, *43*, 164–169. https://doi.org/10.1016/j.lindif.2015.08.020

Borovsky, A., & Elman, J. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of Child Language*, *33(4)*, 759–790. https://doi.org/10.1017/s0305000906007574

Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences*, *45*, 137–143. https://doi.org/10.1016/j.lindif.2015.11.018

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 269–294. https://doi.org/10.1037/0096-3445.104.3.268

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). Elsevier. https://doi.org/10.1016/S0079-7421(03)44005-X

Ferguson, C. A. (1959). Diglossia. *WORD*, *15*(2), 325–340. https://doi.org/10.1080/00437956.1959.11659702

Gettys, S., Imhof, L. A., & Kautz, J. O. (2001). Computer-assisted reading: The effects of glossing format on comprehension and vocabulary retention. *Foreign Language Annals*, *34*(2), 91–99. https://doi.org/10.1111/j.1944-9720.2001.tb02815.x

Godden, D. R., and Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, *93*(1), 103–128. https://doi.org/10.1037/0022-0663.93.1.103

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*(9), 1215–1226. https://doi.org/10.1002/acp.1525

Hulstijn, J. H., & Laufer, B. (2008). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Hunt, A., & Beglar, D. (2002). Current research and practice in teaching vocabulary. In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 258–266). Cambridge University Press. https://doi.org/10.1017/CBO9780511667190.036

Ko, M. H. (2005). Glosses, comprehension, and strategy use. *Reading in a Foreign Language*, *17*(2), 125–143. http:/hdl.handle.net/10125/66786

Kost, C. R., Foss, P., & Lenzini, J. J., Jr. (1999). Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals*, *32*(1), 89–97. https://doi.org/10.1111/j.1944-9720.1999.tb02378.x

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.x

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Kroll, J. F., & Sunderman, G. (2003). Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. In C. J. Doughty, & M. H. Long (Eds.) *The handbook of second language acquisition* (pp. 104–29). Blackwell Publishing. https://doi.org/10.1002/9780470756492.ch5

Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning & Technology*, *3*(2), 58–76. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/7e3aecb4-dbf7-4bf3-9ed6-9a6b9023fd11/content

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. http://dx.doi.org/10.1093/applin/22.1.1

Miyasako, N. (2002). Does text-glossing have any effects on incidental vocabulary learning through reading for Japanese senior high school students? *Language Education & Technology*, *39*, 1–20. https://doi.org/10.24539/let.39.0_1

Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL test battery. *Journal of Neuroscience Methods*, *222*, 250–259. https://doi.org/10.1016/j.jneumeth.2013.10.024

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066661.001.0001

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25

Potter, M. C., So, K.-F., Eckardt, B. V., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, *23*(1), 23–38. https://doi.org/10.1016/S0022-5371(84)90489-4

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Taylor, A. M. (2009). CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, *27*(1), 147–160. https://www.jstor.org/stable/calicojournal.27.1.147

Taylor, A. M. (2010). Glossing is sometimes a distraction: Comments on Chen and Good (2009). *Reading in a Foreign Language*, *22*(2), 353–354. https://files.eric.ed.gov/fulltext/EJ901551.pdf

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*(2), 48–67. http://dx.doi.org/10125/44180

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*(3), 85–101. http://dx.doi.org/10125/44076

Yoshii, M., & Flaitz, J. (2013). Second language incidental vocabulary retention: The effect of text and picture annotation types. *CALICO Journal*, *20*(1), 33–58. https://doi.org/10.1558/CJ.V20I1.33-58

Zarei, A. A., & Mahmoodzadeh, P. (2014). The effect of multimedia glosses on L2 reading comprehension and vocabulary production. *Journal of English Language and Literature*, *1*(1), 1–7.

Appendix A. Language Background Questionnaire

Subject number: _____

This questionnaire is designed to learn about your language history. Information collected will be stored in a secured laboratory. No identifying information will be made available. All answers are strictly confidential. Any information you provide will not be distributed to outside parties. Thank you!

1. Date of experiment: …….
2. Date of birth: …….
3. Place of birth: a. City: .....
   b. State/Province & Country: .....
4. Which Arabic dialect do you speak/learn? .....
5. List all language(s) you know and estimate your ability to speak, understand, read, and write the language(s) on a scale from "1" (i.e., your ability is very poor) to "7" (i.e., your ability is very good).
6. Have you ever had any kind of language learning through technology? YES NO

If "YES", please explain:

7. Have you ever travelled to an Arab country? YES NO
8. Did you play video games as a child? YES NO
9. Do you still play video games? YES NO
10. Do you spend less than 2 hours on your phone daily? YES NO
11. Is Arabic your first "second-language experience"? YES NO

Appendix B. Example Questions from the Pretest

Pre-reading task: Provide the meaning for the vocabulary items that you know. If you do not know, write NS (not sure)

قِصّة

سَاعَدَ

طَلَبَ

هَرَبَ

Appendix C. Examples of Gloss Conditions, Experiment 1

Appendix D. Examples of Target Words and Glosses, Experiments 1 and 2

Exp. 1	Text	Picture	Text+Picture
Arabic word	maʃhuur	ʔarkab	tunaððef
Translation (not in gloss)	famous	I ride; I take transportation	she cleans
Text gloss	maʕruuf	X	naðeef
Translation of text gloss	known	X	clean
Picture gloss	X	(someone getting in bus)	(someone cleaning)

Exp. 2	**Gloss without Root**			**Gloss with Root**		
	Text	Picture	Text+Picture	Text	Picture	Text+Picture
Arabic word	jaħtafel	yatˁtˁsel	raakib	wusˁuulii	naʔeman	Taʃɣeel
Translation (not in gloss)	he celebrates	he calls	Riding or passenger	my arrival	sleeping (literal meaning: sleeper)	starting/operating
Text gloss	ħafla	X	rakeba	ʒeet	X	badaʔ
Translation of text gloss	party	X	to ride	I came	X	He started
Picture gloss	X	(a phone)	(a passenger getting on the bus)	X	(someone sleeping)	(somone starting a car)
Root gloss	X	X	X	wasˁala	Nama	ʃaɣal
Translation of root (not in gloss)	X	X	X	to arrive	to sleep	to start/to operate

Appendix E. Examples of Questions from the Recognition Task

Recognition Task: Choose the English equivalent of the Arabic words below:

Question 1: اعتقلت

1. carried
2. went outside
3. went inside
4. arrested

Question 2: دخل

1. went inside
2. went outside
3. tied
4. untied

Appendix F. Examples of Questions from the Production Task

Production Task

Provide the meaning of the following words in Arabic (write in Arabic):

Question 1: he ran away

Question 2: money

Appendix G. Example of a Root Word Complementing a Picture Gloss, Experiment 2

About the Authors

Juman Al Bukhari is Associate Professor of Arabic and Linguistics at the University of North Georgia, Dahlonega. Besides linguistics and second language acquisition, she is interested in higher education leadership and practice with a focus on internationalization of higher education institutions in the U.S.

E-mail: juman.albuhkari@ung.edu

John A. Dewey is Associate Professor of Psychological Science at the University of North Georgia, Dahlonega. Among other topics, he is broadly interested in applying theories from cognitive psychology to the scholarship of teaching and learning.

E-mail: john.dewey@ung.edu