2018

The Effect of Gloss Type on Learners' Intake of New Words During Reading: Evidence from Eye-tracking

Paul Warren
Victoria University of Wellington

Frank Boers
fboers@uwo.ca

Gina Grimshaw
Victoria University of Wellington

Anna Siyanova-Chanturia
Victoria University of Wellington

THE EFFECT OF GLOSS TYPE ON LEARNERS' INTAKE OF NEW WORDS DURING READING: EVIDENCE FROM EYE-TRACKING

Paul Warren, Frank Boers, Gina Grimshaw and Anna Siyanova-Chanturia

Victoria University of Wellington

Abstract

A reading experiment combining online and off-line data evaluates the effect on second language learners' reading behaviours and lexical uptake of three gloss types designed to clarify word meaning. These are (a) a textual definition, (b) a textual definition accompanied by a picture, and (c) a picture only. We recorded eye movements while intermediate learners of English read a story presented on-screen and containing six glossed pseudowords repeated three times each. Cumulative fixation counts and time spent on the pseudowords predicted post-test performance for form recall and meaning recognition, confirming findings of previous eye-tracking studies of vocabulary acquisition from reading. However, the total visual attention given to pseudowords and glosses was smallest in the condition with picture-only glosses, and yet this condition promoted best retention of word meaning. This suggests that gloss types differentially influence learners' processing of novel words in ways that may elude the quantitative measures of attention captured by eye-tracking.

Keywords: reading; glosses; multimodality; vocabulary; eye movements

Background

A common way of facilitating learners' text comprehension and supporting the intake of new vocabulary is to provide glosses that clarify the meaning of unfamiliar words. Glosses with multimodal content (i.e., both pictorial and textual clarifications) benefit intake of word meaning particularly well, according to post-reading tests in some studies (e.g. Kost, Foss, & Lenzini, 1999; Yoshii & Flaitz, 2002). The reported benefits are often attributed to the advantages of coding information both verbally and nonverbally, in keeping with, for example, Paivio's (1986) dual coding theory and Mayer's (2009) framework of multimedia learning. These accounts imply that adding a pictorial elucidation of word meaning triggers processing of the word that is qualitatively different from processing engendered by a gloss with only a textual clarification.

However, Boers, Warren, Grimshaw, and Siyanova-Chanturia (2017) argue for an alternative, or at least complementary, account for the reported benefits of multimodal glosses, namely that they are less likely to be ignored than text-only glosses and also invite longer processing. However, the extent to which different glosses actually influence learners' engagement with glossed words has not yet been properly investigated, since previous studies only used off-line measures of learning, not online measures of reading behaviour. This is an unfortunate gap in the research, not only because of the theoretical debate, but also because better insights into how gloss types influence reading behaviour may inform the design of pedagogic materials.

The present study is a step towards filling that gap. We recorded eye movements of adult learners of English as a Second Language (ESL) as they read a story containing six pseudowords each accompanied by a textual gloss, a pictorial gloss or a multimodal gloss (i.e., comprising both the picture and the textual gloss). Unannounced tests gauged learners' recall of the form of these target pseudowords and recognition of their meaning. So far, most studies of multimodal glossing have focused on learners' retention of word meaning, but it is worth examining learners' recall of orthographic form as well, because gloss types may differently influence the attention given to form.

Gloss types and vocabulary acquisition from reading

Research on the effects of gloss types is wide-ranging and includes, for example, comparisons of first (L1) and second language (L2) glosses (Ko, 2012) and ways of promoting cognitive engagement through multiple-choice formats (Watanabe, 1997). Here,

however, we focus on studies comparing the effect on L2 vocabulary uptake of glosses with and without pictorial components, mostly in incidental learning conditions.

As an example of research focusing primarily on the uptake of word meaning, Kost et al. (1999) compared three gloss conditions: text-only (L1 translations of test words), picture-only, and multimodal with both text and picture. Learners seeing multimodal glosses outperformed those seeing text-only glosses in a post-test requiring selection of the appropriate picture for each test word. This is unsurprising, since participants in the multimodal condition had seen the pictures used in the test during their reading of the text, while those in the text-only condition had not. This test showed no significant difference between multimodal and picture-only conditions. A second post-test involved matching test words with L1 translations. Unsurprisingly again, performance was better if participants had seen the L1 text glosses than if they had only seen pictures. However, the multimodal condition produced significantly higher scores than either of the other conditions. So, seeing both the illustration and the L1 translation in the gloss appeared to produce better retention of word meaning than seeing just the translation.

Yoshii and Flaitz (2002) conducted a replication of Kost et al. (1999) in a multimedia context. In a picture-recognition post-test, participants in the multimodal condition again outperformed the text-only group, with no significant difference from the picture-only condition. As in Kost et al. (1999), picture-only glosses were relatively unhelpful when participants were asked to either supply or recognize definitions of the words. Overall, the trend was again for multimodal glosses to result in greater learning than either single gloss condition.

In another computer-based multimedia presentation of reading materials, Chun and Plass (1996), Plass, Chun, Mayer, and Leutner (1998) and Akbulut (2007) accompanied target words with textual and visual annotations. Participants indicated which annotation(s) to consult by clicking on hyperlinks. More correct post-test responses were given if both text and picture annotations had been consulted than just text annotations. Similar results were found by Jones and Plass (2002), using a reading-while-listening task where the participants could pause the recording and choose to click links to textual and/or pictorial word annotations. However, the results of these experiments may in part be due to the number rather than type of annotations consulted for each target word (Boers, Warren, Grimshaw, et al., 2017).

While these studies show benefits of multimodal glosses for the retention of word meaning, there are some exceptions: Neither Acha (2009) nor Boers, Warren, He, and Deconinck (2017) found evidence supporting multimodal over text-only glosses.

Eye-tracking studies with a focus on vocabulary

The studies summarised above, like most studies of vocabulary uptake from reading, employed post-reading tests, which gauge the outcome of the reading process but have little to say about that process itself. To better understand the latter, in both L1 and L2, researchers have studied eye movements during reading. Eye-movement research (e.g., Rayner, 1998, 2009) uses measures such as visual fixation on a word as proxies for attention (see Godfroid and Schmidtke (2013) for a discussion of the relationship between visual fixation and the concepts of attention and awareness).

To date, few studies have researched eye-movement patterns in the context of incidental vocabulary acquisition during reading. These have shown that unfamiliar words attract more attention than familiar ones, that attention paid to novel words predicts their retention in memory, and that the increased attention declines over multiple encounters. We review here some of that evidence.

In an early study, Chaffin, Morris, and Seely (2001) investigated how readers establish the meanings of new words during silent L1 reading. Sentence pairs comprised a sentence containing the target word and a second sentence containing a related word. The first sentence was either neutral or highly informative concerning the target. Reading times for this first sentence depended on target familiarity (e.g. *guitar* vs. *zither* vs. the pseudoword *asdor*) and on the informativeness of the context. Crucially, readers spent more time on the related word in the second sentence (e.g. in this case *instrument*) when the target word was unknown and the context was uninformative (neutral). This suggests that readers successfully identify and pay attention to the portion of the text that is relevant for inferring the meaning of an unknown word, which, in the case of an otherwise uninformative context, was the related word. As the authors acknowledge, however, they collected no direct evidence that learning of the novel word's meaning had taken place (though they cite supporting evidence from a previous study using the same materials and a direct assessment of understanding).

In a study of the effects of word familiarity in silent L1 reading, Williams and Morris (2004) measured both visual fixation patterns and learning outcomes. Both initial and second-pass reading time (i.e. re-reading) indicated that unfamiliar words received more attention. Post-

reading tests showed better meaning retention for many of these unfamiliar words. Interestingly, initial reading time was shorter and second-pass reading time longer on unfamiliar words that received correct post-test responses than on those that received incorrect responses. Since the preceding sentence context was not highly informative for interpreting the unknown word, the authors interpret this behavioural pattern to show that readers do not dwell on novel words if they have little information to aid their interpretation, but will return to them (as reflected in second-pass reading times) if the following context gives them something to work with.

More recently, Brusnighan and Folk (2012) investigated the role of morphological and contextual information in incidental vocabulary acquisition during L1 reading. Eye movements were measured as participants read English compounds that were either familiar and transparent (*milkshake*), novel and transparent (*drinkblend*), familiar and opaque (*cocktail*) or novel and opaque (*deskdoor*), each occurring in either neutral or informative sentence contexts. Post-tests showed that their meanings could be retained after a single encounter, even for novel forms. As expected, novel items received longer reading times than familiar items. In addition, morphologically transparent novel items embedded in informative contexts, i.e. when both morphology and context cued the meaning of the item, had shorter re-reading times, indicating a processing advantage.

The relationship between eye movements (as a proxy for attention) and pseudoword learning in short L2 passages was explored by Godfroid, Boers, and Housen (2013),. Post-reading recognition was measured by presenting learners with the sentence in which a pseudoword had been originally encountered, with the pseudoword replaced by a dotted line, and requiring them to select the appropriate item from 18 candidate items, including the pseudoword and 11 other pseudowords used in the experiment. There was a positive correlation between fixation times on pseudowords and their recognition accuracy.

A range of post-tests was employed by Pellicer-Sánchez (2016), including immediate and delayed tests of the acquisition of both word meaning and word form. Both L1 and proficient L2 speakers read stories for comprehension. The study also considered the effect of repetition, with novel words appearing multiple times in contexts deemed helpful for inferring meaning. The two reader groups differed in acquisition rate, but not in terms of outcome. After eight exposures to the novel words, L2 readers recognised their form and meaning in multiple-choice tests with 86% and 75% accuracy respectively, and their success

in supplying word meaning was 61%. The eye-movement data showed decreases in fixation counts and durations after three or four encounters. By the eighth encounter they were read similarly to known words. Again, greater total reading times on the novel words predicted better learning.

In a similar study investigating multiple encounters with novel words during story-reading, Mohamed (2017) also found a gradual decrease in fixation durations as advanced L2 readers became more familiar with the words. The novel words varied in the number of times (1-30) they occurred in the story. Once again, total reading duration predicted learners' performance on post-tests of word meaning and form, and was a stronger predictor than the number of instances of the novel words.

Finally, Elgort, Brysbaert, Stevens, and Van Assche (2017) studied L2 readers' eye movements as they read an expository text (chapters from a general-academic book). Again, unfamiliar words occurred with varying frequencies (8-64 occurrences). Time spent on the words decreased, most markedly over the first ten encounters. Nevertheless, even after as many as 40 encounters noticeable differences in fixations and reading times remained between target words and familiar control items. Participants' mean success in supplying target word meanings in a post-test was 34%. Both of these findings contrast with those of Pellicer-Sánchez (2016), perhaps because of the helpful contextual cues that accompanied the target words in the latter.

Likely owing to the complex study design (e.g., the reading was spread over two days, introducing a longer time interval between word encounters and the post-test), Elgort et al. (2017) did not report an association between total reading times and post-test performance. Still, the other studies reviewed above suggest a positive relationship between fixation times and learning outcomes. In addition, attention devoted to novel words during reading can depend on their transparency and on contextual support. While we might expect glosses to be an important source of supporting evidence for novel words, none of the reported eye-movement studies included the impact of glosses on reading behaviours or learning outcomes. In the study reported below, we extend previous reading time studies to include time spent fixating both the novel words and the glosses, allowing us to examine the effect of attention to glosses on reading behaviour and on learning outcomes. This may, for example, help to (re-)interpret the advantage (if any) of multimodal over single-mode glosses in terms of quantity rather than quality of processing.

The current study

We have highlighted two research strands concerning vocabulary acquisition during reading. One shows the impact of gloss types on learning outcomes, particularly meaning retention. The other uses eye movements to demonstrate a positive relationship between attention paid to novel words and memory for these words, in situations where readers are usually left to their own devices to discover their meanings. To our knowledge, there is no published research which examines whether the positive relationship between fixation time and uptake in memory helps to account for the superiority (if any) of multimodal glosses over single-mode ones for word learning.

In the current study, learners were exposed to novel words (pseudowords) in a story context. The pseudowords were accompanied by textual, pictorial or multimodal (text and picture) marginal glosses. While our learners read the story, their eye movements were recorded, providing a record of attention to pseudowords and glosses. Learning was measured by post-tests for both pseudoword meaning and form.

Across the story, readers were exposed to three instances of the pseudowords (along with one in the gloss). This adds to the ecological validity of including glosses, since these are more likely to be provided for important novel words, which are in turn more likely to recur in a text.

Research questions

We address the following general questions:

1) Do multimodal marginal glosses (with both textual and pictorial clarifications of word meaning) help L2 learners achieve better scores than single-mode marginal glosses (with only a textual or pictorial clarification) on post-reading tests concerning word form as well as word meaning?

2) How much attention (as measured by eye-tracking) is devoted to instances of the novel words and to their glosses, and how is this attention affected by the gloss type?

3) To what extent does the amount of attention paid to instances of the novel words and to their glosses predict performance in post-reading tests of word form and meaning?

Participants

Our readers were 52 adult high-intermediate ESL learners (30 females, 22 males) enrolled in a general English proficiency programme at Victoria University of Wellington.[1] They volunteered after reading an information sheet which explained the study was about reading in a second language. The precise aim of the study was not specified. The study had approval from the University's Human Ethics Committee (approval #20143). Learners were randomly assigned to one of three treatment conditions (see below) and received a NZ$20 supermarket voucher in return for their time.

Materials and procedure

The reading text was based on a local news story, was approximately 900 words long, and incorporated six pseudowords. While we recognize that six target items represents a rather small number of experimental tokens, we were keen to ensure that the text remained intelligible and that our readers would read with a primary focus on the story's content without being distracted by excessive information in the margin. Apart from the six pseudowords, two real English nouns were glossed, to reduce the likelihood of participants becoming suspicious of the nature of the target pseudowords. The proportion of eight glossed words in a 900-word text is similar to a recent study by Khezrlou, Ellis, and Sadeghi (2017).

Each pseudoword occurred three times in the text body, twice on the same page as the gloss, and once on the following page. On its first occurrence in the story each pseudoword was presented in boldface – typographic enhancement being a typical means in glossing interventions to indicate that an annotation about the word is available – and was accompanied in the right-hand margin by a gloss with one of the following forms:

a) Pseudoword (bolded) followed by a textual definition ('text-only');
b) Pseudoword (bolded) accompanied by a picture above it ('picture-only');
c) Pseudoword (bolded) followed by a textual definition and accompanied by a picture above it ('multimodal').

Three versions of the text were created, one with each type of gloss. Fifteen participants completed the text-only condition, 19 the picture-only condition, and 18 the multimodal condition.

[1] Ten further participants were excluded because of unreliable eye-tracking data.

The second occurrence of each pseudoword in the body of the text was on the same page as the first, an average of 48 words / 6 lines later. The third instance was on the following page. The target word was the first content word in the textual definitions. Appendix A gives examples of the three conditions for the pseudoword *perchant*.

The forms of the glossed pseudowords (*panipline, perchant, hangles, dasters, bandilon,* and *stavener*) were borrowed or adapted from Godfroid et al. (2013) and Godfroid and Schmidtke (2013). Phonological and orthographic plausibility in English was verified by four native speakers. All pseudowords replaced words with a concrete meaning, to enable easy pictorial elucidation. The pictures were colour photographs selected from freely available internet materials. To ascertain that the pictures were unambiguous, nine PhD students read the text with picture-only glosses and subsequently named or described the referent they thought each picture represented. No evidence of picture ambiguity emerged.

The text was distributed over eight screens, with each introducing a new glossed word. Participants pressed the space bar to move to the next screen when ready. They could not return to previous screens. Participants were told that a quiz about the text would follow, but not that it would involve vocabulary. To a degree, then, the current study examines incidental vocabulary learning, although it should be noted that our reading conditions are different from incidental learning conditions in many previous studies since the presence of glosses potentially encourages learners to focus on the form, thus making the form more salient. The quiz (henceforth post-test) was administered using E-Prime (Psychology Software Tools, 2012) immediately after the reading activity, and comprised two parts, In the first part, participants were asked five multiple-choice content questions (each with one correct option and three foils) and eight word recall questions. Each question was presented on a separate screen. The content questions concerned text passages that did not require comprehension of the pseudowords. The word recall questions concerned the glossed words: six pseudowords and two existing words. For each word, three recall prompts were presented together on the screen: (a) the sentence in which the target word was first encountered in the text, with a gap instead of the word; (b) the definition from the textual gloss; and (c) the picture from the pictorial gloss. Participants were asked to type as much as they could remember of the form of the word. An example of the recall test for word-form is provided in Appendix B.

The second part of the post-test was a matching task. Each pseudoword was presented on a separate screen and participants were asked to identify its corresponding meaning from

among 11 options. These consisted of the pictures and definitions (as given in the textual glosses) of the six pseudowords and two real words glossed in the text and of three additional real words that occurred in the text. An example of the matching task used to test meaning retention is provided in Appendix C.

The prompts used in both tests were the same for all treatment conditions. It could be argued that confronting some participants with pictorial representations that they had not seen during their reading activity may have created an additional processing task at the test stage. Likewise, giving participants textual definitions which they had not previously seen may have induced extra processing. Note that the non-congruency between gloss condition and test prompts that occurred in the context of previous studies (e.g., Kost et al., 1999) was where participants in some conditions were required to rely exclusively on test prompts they had not seen during reading. Our primary rationale for using identical tests for each treatment group, including a combination of prompts, was that we could be sure that they at least included the meaning representation available during reading. In addition, using a tailor-made test for each treatment condition would arguably have introduced a confounding factor, where different outcomes might have been attributed to the test condition rather than to the reading condition. Finally, no time pressure was imposed in the post-tests, and so, even though the presence of new stimuli probably invited extra processing in two of the treatment groups, this is unlikely to have had a detrimental effect on the accuracy of responses due to an excess of information.[2] The reading and testing procedures together took about 20 minutes.

Participants were randomly assigned to one of the three conditions. They came from five different language classes, to which they had been assigned based on an English proficiency test. Since we did not have human ethics approval to access participants' proficiency test scores, we used their class level as a proxy for proficiency. There was no significant difference in the class level of participants in the three conditions (by chi-square test, $p = .98$). In addition, there was no significant difference between groups in their scores on the general comprehension questions about text content (Kruskal-Wallis $\chi^2(2) = .24$; $p = .89$).

[2] E-prime did record response times, but these are not reported here. It is worth mentioning, however, that responses were slower when material in the test prompts had not previously been seen (i.e., the pictorial material when the glosses were text-only and the definitions when the glosses were picture-only). This lends some support to the thesis that the test mode was optimally congruent with the multimodal gloss condition but less so with the other two conditions.

Apparatus

Participants were tested individually. A forehead and chin rest maintained a viewing distance of 57 cm to a 21-inch computer monitor, which presented pages of text at a resolution of 1024 x 768 pixels and a refresh rate of 60 Hz. At this distance and resolution, text presentation averaged three letters per degree of visual angle. Eye movements and positions were measured with an EyeLink 1000 Tower Mount Head Supported system (SR Research Ltd., Ontario, Canada). A video-based infrared camera measured corneal reflection and the dark pupil of the right eye via an infrared reflective mirror. Positions were sampled at 1000 Hz, and at a spatial resolution of 0.01 degrees of visual angle. Experiment presentation was controlled using SR Research Experiment Builder Version 1.10.165 (2011).

Fixation data were tabulated for a number of pre-defined interest areas, including the three presentations of each pseudoword in the text, the pseudoword in the marginal gloss, the picture in the gloss (for conditions including a picture), and the textual gloss (for conditions that included this). Fixations were extracted using Eyelink Data Viewer Version 1.11.9000 (2007). A research assistant blind to our hypotheses conducted drift correction using Data Viewer's semi-automatic algorithm. The procedure involved placing the first fixation on each line of text manually, followed by automatic correction of subsequent fixations on the line. After drift correction, fixations were cleaned in a 4 step-process by which brief fixations within a radius of 0.5 degrees were merged. After merging, fixations of less than 140 ms or greater than 800 ms were deleted. This procedure resulted in a loss of 8.46% of fixations, which is within typical levels, and importantly did not differ by treatment condition.

Attention measures

We examined three eye-tracking measures: First Fixation Duration (FFD), Total Fixation Counts (TFC), and Total Reading Time (TRT). These serve as proxies for the amount of attention paid by participants to each of the three occurrences of the pseudowords in the body of the text, and to the marginal glosses (and their components: the pseudoword itself, its textual definition and/or its picture). In terms of the relationships between eye-tracking measures and cognitive processing discussed by Johnson and Mayer (2012: 181-2), we are considering measures of attentional focus on pseudowords and glosses, rather than measures of integration of pseudowords and glosses, which these authors argue are best measured by considering transitions (i.e. saccades) between the text and the glosses (see also Mason,

Pluchino, & Tornatora, 2016, who focus in particular on transitions between text and illustration during second-pass reading).

An additional measure is whether any attention is paid to the interest area, or whether instead it appears to be ignored. Reading studies have shown that if a word is predictable in its context, then it is frequently skipped (Rayner, Slattery, Drieghe, & Liversedge, 2011), with as many as a third of all words skipped in initial reading (Rayner, 1998, 2009). In our study, fixation likelihoods for the pseudowords were at or close to ceiling. This is not unexpected – participants were non-native speakers who mostly read slowly through the text, and the pseudowords themselves were both unknown and appeared in the types of sentence position expected of content words that are highly likely to be fixated (Rayner, 1998: 375 reports that native English speakers fixate content words approximately 85% of the time, and function words 35% of the time).

Nevertheless, the fact that words (and glosses) can be skipped is problematic for the analysis of FFDs and TRTs, and in particular for whether averages of these measures should be based on just those cases where there are measurable fixations, or should also include zero values. Rayner (1998:376-8) points out that words can be processed during reading even when they are not fixated since some features such as overall length and shape will become available to the reader (via parafoveal preview) during fixation on earlier words. He discusses a number of solutions to the problem of what to do with non-fixated words. Many of these solutions are less appropriate for our study, where we are interested not just in fixations on individual words during reading, but also in reading behaviours over larger areas of interest such as the entire marginal gloss, or the picture in a gloss. In this context it is important to mention that the glosses were more frequently ignored, i.e., not fixated at all, than the pseudowords. This was particularly striking in the case of the text-only glosses, which were ignored 20.0% of the time. This compares to 11.4% and 9.3% for the picture-only and multimodal glosses, respectively. If text-only glosses were ignored so much more often than the other two gloss types, this inevitably has implications for whether we should exclude missing fixations from our comparisons of average Fixation Counts and average TRTs. We chose to include zero-fixations, however, because we are interested in how (if at all) different gloss types influence reading behaviour. If some gloss types are more likely to be ignored, then that is a pedagogically pertinent finding.

Our analysis of TRTs, as a measure of total attention given to an item, includes two components: skip rate (i.e. the probability that the item receives no fixation) and the total time spent fixating the item, conditional upon skip rate. For the analysis of FFDs, however, we follow Pollatsek, Reichle, and Rayner (2003: 371) and Murray (1998, p190: fn4) in treating any skipped instances of the item in question as missing values. Therefore, the average FFDs reported below should be interpreted as indicating the amount of attention given to an item on first seeing it, given that it receives any attention at all.

Data analysis

In the form recall test, participants seldom supplied a fully correct target word. Recall attempts usually resulted in partially correct responses (e.g., just the first letter or syllable). Two blind judges independently scored each response on an 11-point scale from 0.0 (no form recall attempt or a completely incorrect response) to 1.0 (completely correct recall). Intermediate scores included 0.1 when only the first letter was given, 0.3 for the incomplete response *pan* (for *panipline*) and 0.8 for the nearly accurate response *banlion* (for *bandilon*). Since interrater agreement was high (r = .98), we used the average score of the two judges in our analysis. The responses on the meaning recognition test were scored in a binary fashion (i.e., either correct or wrong).

The results section below presents analyses of how readers directed their attention (with respect to the pseudowords and the glosses), of the overall impact on form recall and meaning recognition of different gloss types, and of the relationship of attention to form recall and meaning recognition. For most of the statistical analyses of the attention measures and the relationship of these to form and meaning retention, we ran mixed effects models using *lmer* from the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in *R* (R Core Team, 2014). Linear models were used in all cases except for when meaning recognition (a binary variable) was the dependent variable, where logistic models were used. When the dependent variable was TFCs, the linear model assumed the Poisson distribution that is appropriate for count data. Attention measures that involved durations (FFD and TRT) were log-transformed prior to statistical analysis to provide a better fit to a normal distribution. For TRTs we used a zero-inflation approach to mixed effects modelling using *glmmTMB* (Magnusson et al., 2017). This permitted a two-part analysis: a logistic model predicted the likelihood of skipping the item in question and a linear model predicted reading time (based on log values), conditional on skip rate.

Barr, Levy, Scheepers, and Tily (2013) recommend fitting a maximal random effects structure, with appropriate by-participant and by-items slopes for fixed effects. However, in the our data this typically resulted in overfitting and in model non-convergence, which is not surprising given the small number of items (pseudowords). For each model reported below we therefore obtained an optimal random effects structure following the procedure outlined by Bates, Kliegl, Vasishth, and Baayen (2015). The significance level of fixed effect predictors was assessed via model comparison using likelihood ratio tests with the *mixed* function in the *afex* library (Singmann, Bolker, & Westfall, 2015). Note that *mixed* uses sum contrast coding, comparing each level of a factor to the average of the other levels. Post-hoc comparisons, with Tukey p-value correction for multiple contrasts, were performed on model outputs through least-squares means using *lsmeans* (Lenth, 2016). Further details of statistical model design, including the optimal random effects structures, are given in relevant part of the results section.

Results

Attention

In this section we present analyses of the attention paid to the pseudowords and their glosses during participants' reading of the text. Since the glosses differed by condition, separate analyses involved attention paid to the pseudowords and to the glosses. Recall that each pseudoword appeared three times in the text and once in the gloss. FFDs, TFCs and TRTs were the dependent variables in separate mixed effects regression analyses. Fixed effects were Instance (the three instances of the pseudoword in the text and the one instance in the gloss), Condition (multimodal, picture-only, text-only), the interaction of these factors, as well as pseudoword Length (in characters) for analyses involving pseudowords (since word length is widely recognised to have an effect on fixations, e.g., Rayner, 1998, 2009), and the Class level of the participant (as a proxy for proficiency). The factor Instance was treated as categorical since there was no fixed order in which the four occurrences of the pseudoword had to be inspected, save that the third instance in the text was always seen last since it occurred on a following screen page. A descriptive summary of results is provided in Table 1.

<Table 1 about here>

First Fixation Durations

The optimal random effects structure for the analysis of FFDs on the pseudowords themselves included just intercepts for participants and pseudowords. The significant effects returned by model comparison were Instance ($\chi^2(3) = 27.15$, $p < .0001$) and the interaction of Instance and Condition ($\chi^2(6) = 17.09$, $p < .01$). The Instance effect reflects significantly shorter FFDs on the pseudoword in the gloss – despite it being in bold face – than on each of the instances in the text ($p < .02$ in least-squares means comparisons). FFDs for the instances in the text did not differ from one another. As can be seen from Figure 1, the interaction reflects differences in the patterns of FFDs across the four instances in the different conditions. Post-hoc pairwise comparisons showed no significant differences between any of the instances in the picture-only condition. In the multimodal condition, FFDs on the first instance were significantly greater than those on the instance in the gloss ($t = 4.45$, $p < .001$) and on the third instance ($t = 2.65$, $p < .05$), and longer on the second instance than on the instance in the gloss ($t = 3.20$, $p < .01$). In the text-only condition, both the first instance and the instance in the gloss had shorter fixations than the second instance ($t = -2.79$, $p < .05$ and $t = -3.76$, $p < .01$), and the instance in the gloss had shorter fixations than the third instance ($t = -2.68$, $p < .05$). There is thus little evidence here that typographic enhancement (bold face) of the first instance of the word in the text and of its iteration in the gloss triggered consistently longer FFDs. This is surprising, given that typographic enhancement had an attention-directing effect in other eye-tracking studies, such as Winke (2013) and Choi (2017) – but see below for Fixation Count data that accords better with expectations.

<Insert Figure 1 about here>

FFDs in the definition area were compared in the multimodal and text-only conditions, since there was no definition in the picture-only condition. Similarly, the picture area could only be compared in the multimodal and picture-only conditions. In both analyses, the mixed effects models tested for Condition and Class as fixed effects, and the random effects structure included intercepts for participants and pseudowords, and random slopes for Condition across pseudowords. The only significant effect for the definition area was Class, with longer first fixations for participants in the more advanced classes ($\chi^2(1) = 4.21$, $p < .05$). Neither Condition nor Class showed significant effects for first fixations in the picture area.

Fixation Counts

We turn now to the analysis of TFCs, starting with the pseudowords themselves. The optimal random effects structure included intercepts for both participants and pseudowords, as well as participant and pseudoword slopes for Instance. The analysis returned significant effects of Instance ($\chi^2(3) = 18.374$, $p < .001$) and Length (longer pseudowords had more fixations: $\chi^2(1) = 5.97$, $p < .05$). The model predicted more fixations on the first instance of the pseudoword than on the others (each of which was significantly different from the first instance by least-squares means at $p < .01$, and did not differ from one another). Condition had no effect on the number of fixations on the pseudowords.

For TFCs in the definition and picture areas, the analyses again tested for Condition and Class, with random intercepts for participants and pseudowords and random slopes for Condition across pseudowords. Neither fixed effect was significant in the analysis for the definition area. For the picture area there was a significant effect of Condition ($\chi^2(1) = 5.39$, $p < .05$), with more fixations in the picture area in the picture-only condition than in the multimodal condition. Class was not significant.

A comparison of the aggregated TFCs in all areas of interest concerning the pseudowords, i.e., their instances in the text plus their associated gloss components, shows that numerically the multimodal (17.5) and text-only gloss conditions (16.8) had more fixations than the picture-only gloss condition (13.8), but this difference was not significant ($p = .15$).

Total Reading Time

The optimal random effects structure for the zero-inflation analysis of TRT on pseudowords included just random intercepts for both participants and pseudowords. There were significant overall effects of Instance ($\chi^2(6) = 213.00$, $p < .0001$) and Length ($\chi^2(2) = 6.63$, $p < .05$). Consideration of the two components of the optimal model shows that the logistic analysis of skip rate produced a significant effect of Instance, but not of Length, while the linear analysis of TRT conditional on skip rate was significant for both Instance and Length. The Length effect was that longer pseudowords had greater TRTs. To assess significance of the differences between scores for each instance, models were run with different instances as the baseline value against which the other instances were compared (by z-test). There were significantly fewer skips of the first instance of the pseudoword than any other instance and significantly more skips of the pseudoword in the gloss than for any of the other instances ($p < .001$ in each case). Skips for the second and third instance did not differ from one another.

TRT was significantly greater for the first instance than for any of the others, and significantly shorter for the second instance than for any of the others ($p < .01$ in each case). TRTs for the third instance and for the instance in the gloss did not differ from one another.

Analysis of the entire definition area (excluding the picture-only condition, which did not have a textual definition) included Condition and Class as fixed effects. The optimal random effects structure included only intercepts for participants and pseudowords. There were no significant effects. The analysis for the picture area (excluding the text-only condition) similarly included Condition and Class as fixed effects, but this time the optimal random effects structure included by-pseudoword slopes for Condition as well as intercepts for both participants and pseudowords. There were no significant effects in the logistic part of the model, i.e. skip rates did not differ by Condition. TRTs did, however, differ significantly by Condition ($z = 2.5$, $p < .05$). The average total time spent looking at the picture was longer in the picture-only condition than in the multimodal condition. The fact that readers spent more time inspecting the picture in the former condition is not unexpected as the picture was the only information available to them in the gloss area to figure out the word's meaning.

We also ran regression models for overall Fixation Counts and TRT on the set of four instances of each pseudoword pooled together, with Condition, Length and Class as fixed effect predictors, along with random intercepts for participants and pseudowords and by-pseudoword slopes for Condition. The analysis of summed Fixation Counts returned no significant effects. When all four instances are pooled in this way, skip rates are virtually nil, meaning that a zero-inflation analysis of TRT was not possible. The analysis of TRT is therefore of the time spent reading the pseudowords, summed over the four instances, without consideration of skip rate. The only significant effect was of pseudoword length ($\chi^2(1) = 5.47$, $p < .05$), with longer reading times for the longer pseudowords.

Finally, we also added up the TRTs per pseudoword and its associated gloss components. Altogether, the words and their glosses attracted the greatest average TRT in the multimodal condition (4659 ms), followed by the text-only gloss condition (4502 ms). The mean value was markedly shorter in the picture-only gloss condition (3751 ms). This is perhaps not surprising, as there was no textual definition for the participants to 'take in', and it stands to reason that glancing at an elucidating illustration takes less time than reading a definition. It is worth noting, however, that the difference in cumulative TRTs between the three gloss conditions nevertheless falls short of significance ($p = .20$).

Summary

Altogether, the attentional measures lend only modest support to the thesis that gloss types differently affect fixation behaviour. Less time tended to be spent on the definition when a picture was also available in the gloss – which could be interpreted either as competition for attention or as a reflection that the picture facilitated fast processing of definitional content – but this difference did not reach statistical significance. Conversely, when no definition was available in the gloss, the picture received more attention, and this difference did reach significance.

As to the three instances of each target word in the text, the first instance attracted most fixations and longer TRTs. This was to be expected, not only because it was the first encounter but also because it was typographically enhanced. Interestingly, the third instance, which appeared on the next screen and without a clarifying gloss, also tended to attract comparatively more attention. However, the gloss conditions did not appear to differently affect the amount of attention given by the readers to the three pseudoword instances in the body of the text.

Post-reading word-form recall and meaning recognition

Mean scores in the form recall and meaning recognition tests are shown in Table 2. In the form-recall test the picture-only condition yielded the best results, although the effect of Condition was not significant ($\chi^2(2) = 3.53$, $p = .17$). Given the overall lack of indications from the online reading measures that the three gloss conditions directed attention to the pseudowords to different degrees, this is actually unsurprising. We did find a significant effect of Class ($\chi^2(1) = 4.86$, $p < .05$), showing that the more advanced learners were more accurate in their recall of the form of the pseudowords.

<Insert Table 2 about here>

The scores in the meaning recognition test did differ significantly by Condition ($\chi^2(2) = 6.42$, $p < .05$). Average scores in the text-only condition were significantly lower than those in the picture-only condition. Scores in the multimodal condition lay between these two, but were not significantly different from either. In other words, there was a general benefit for meaning retention of having the picture present in the gloss, but this is less marked for the multimodal condition than for the picture-only condition. Class was not significant as a predictor of meaning recognition.

If fixations and reading times are predictive of learning – as has been attested in several studies (see Background above), then it is intriguing that the picture-only glosses appeared to foster the best retention of form-meaning connections, given that we see no evidence that pictures enhance uptake by affecting the attention paid to words or the glosses. Indeed, the aggregated TFCs and TRTs were actually numerically (though not significantly) lowest in the picture-only condition.

Effects of attention paid to the pseudowords

To determine whether the amount of attention paid to the pseudowords in the text influenced their uptake, analyses were carried out in which scores in the tests for form recall and meaning recognition were predicted by each of the three measures derived from our eye-tracking data. Separate series of mixed effects regression models were run with form recall and meaning recognition as dependent variables. In each series, separate models included each of the reading measures (FFD, TFC, or TRT) as a fixed effect predictor. Pseudoword Length and the Class level of the participants (as a proxy for proficiency) were also included as predictors. Initial models considered the interaction of the reading measure with the Instance of the pseudoword (first, second or third in the text, or the one in the gloss). The presence of such an interaction would indicate a differential effect on the dependent variable (form recall or meaning recognition) of attention paid to each instance of the pseudoword. When there was such an interaction, it was further explored through separate models for each Instance. Our models also included Condition (text-only, picture-only and multimodal) and its interaction with the reading measure, as well as the three-way interaction of Condition, Instance and the reading measure (to assess the possible impact of differences in the attention paid to the various instances of the pseudoword that might result from differences between the gloss conditions).

Form recall

The optimal random effects structure for the model testing the effects on form recall of FFD included random intercepts for participant and pseudoword and by-pseudoword slopes for Condition. The model returned a significant effect of Instance ($\chi^2(3) = 8.36$, $p < .05$) and an interaction of this with FFD ($\chi^2(3) = 8.42$, $p < .05$). The only other significant effect was of Class – as reported earlier, participants in the more advanced classes showed better form recall ($\chi^2(1) = 4.79$, $p < .05$). The interaction of Instance with FFD is shown in Figure 2. Further exploration of this interaction considered the effect on form recall of FFD for each instance separately, with Condition, Length and Class included as before. For each of the

three instances of the pseudowords in the text the only significant effect was the participants' Class level (in each case $p < .05$, as in the overall analysis). For the instance of the pseudoword in the gloss, there was additionally a significant positive effect of FFD ($\chi^2(1) = 8.09$, $p < .01$).

<Insert Figure 2 about here>

For our model including the effects of Fixation Count in predicting form recall, the optimal random effects structure included only intercepts for participants and pseudowords. The only significant effects were Class (better recall for higher classes, $\chi^2(1) = 5.37$, $p < .05$) and TFC ($\chi^2(1) = 6.62$, $p < .05$). A larger number of fixations on the pseudowords had an overall facilitative effect on form recall, and this was not affected by gloss Condition nor by which Instance of the pseudoword was fixated.

The model predicting the effect of TRT included random intercepts for participants and pseudowords and by-pseudoword slopes for Condition. The only significant predictor was participants' Class level ($\chi^2(1) = 5.33$, $p < .05$). The TRT spent on each pseudoword had no effect on the recall of form, neither as a simple effect nor in interaction with either Condition or Instance.

The impact of attention paid to the pseudowords across the entire reading passage was assessed by summing TFCs and TRTs for all four instances of each pseudoword. Summed TFCs were used as a predictor in a model that also included Condition (and its interaction with TFC) and Class, along with an optimal random effects structure of participant and pseudoword intercepts and by-pseudoword slopes for Condition. The model returned Class ($\chi^2(1) = 7.08$, $p < .01$) and TFC ($\chi^2(1) = 12.50$, $p < .001$) as significant predictors of form recall, with no other effects. A parallel process was followed for summed TRTs, except that by-pseudoword slopes for Condition had to be dropped before the model would converge. Summed TRT significantly predicted form recall ($\chi^2(1) = 4.60$, $p < .05$). There were no other significant effects.

These results show that form recall increases with the total number of fixations on the pseudowords, and the total time spent looking at them. These appear to be general findings not specific to individual instances of pseudowords. However, we also found that the impact of initial attention paid to pseudowords (as measured by FFD) was limited to the instance in the gloss. None of these effects were influenced by gloss Condition.

Meaning recognition

As with form recall, the analysis involved separate models in which the scores for meaning recognition were predicted by each of the three different reading measures in interaction with Instance, as well as by Condition, the interaction of Condition with the reading measure and the three-way interaction of Condition, Instance and the reading measure. The Length of the pseudoword and the Class level of the participant were also included as fixed effects. The random effects structure for the analyses of FFD and of TRT included random intercepts only. For TFCs, it also included by-pseudoword slopes for Condition. None of the models returned any significant effects.

While there was no impact on meaning recognition of attention paid to individual instances of the pseudowords, further models including summed TFCs and TRTs as predictors showed a different picture. The first model showed a significant effect of TFC summed across all four instances ($\chi^2(1) = 7.62$, $p < .01$), as well as a significant effect of the participants' Class level (higher proficiency participants showed better meaning recognition; $\chi^2(1) = 3.86$, $p < .05$). The second model showed a significant effect of summed TRT ($\chi^2(1) = 4.60$, $p < .05$). In other words, a greater cumulative exposure to the pseudowords bears a positive relationship to meaning recognition.

Effects of attention paid to the information provided by the glosses

We turn now to the effects on form recall and meaning recognition of attention paid to areas of the marginal gloss. Analyses of reading measures for the picture involve comparisons of the picture-only and multimodal conditions, while those for the definition involve the text-only and multimodal conditions. The optimal random effects structure in both models included by-pseudoword slopes for Condition as well as intercepts for participants and pseudowords. Neither analysis returned any significant effects. It appears that the amount of attention paid to the picture and to the definition in the gloss area had no effect on either the recall of the form or recognition of the meaning of the pseudowords.

Discussion

The first general question we set out to answer was whether multimodal glosses are superior to single-mode glosses as regards vocabulary uptake. Our post-test data yield mixed findings. According to the descriptive statistics (Table 2), multimodal glosses appeared to promote better uptake of both form and meaning than text-only glosses, but these differences fell short of statistical significance. Instead, it was the second single-mode gloss type, i.e., the picture-

only gloss, that appeared to promote the best uptake of both form and meaning in the present study, although scores in this condition were significantly different only from those in the text-only condition, and only for meaning recognition. In sum, the test results yield no compelling evidence in favour of multimodal glosses, but are favourable of picture-only glosses – at least in the case of word meanings that are depicted in an unambiguous way. These results differ from several previous studies (see Background) which claimed superiority of multimodal glosses over both text-only and picture-only glosses.

The second research question was whether the gloss types influence the amount of attention given to the target words and/or the glosses themselves. Altogether, we found very little evidence that the three gloss conditions affected the distribution of attention. The one significant difference which emerged was the increased time given by participants to the picture in the glosses when this was the only elucidation presented to them, in comparison to the multimodal glosses where participants appeared to divide their attention between the picture and the textual clarification. The total, cumulative amount of time that participants fixated the words and their glosses was greatest in the multimodal condition, but did not differ significantly among the three conditions.

Regardless of gloss condition, the amount of attention decreased quite dramatically from first to subsequent instances of the target words in the text. Pellicer-Sánchez (2016), Mohamed (2017) and Elgort et al. (2017) also found decreasing reading times on repeated occurrences as the novel words gradually became more familiar, but the sharp decline observed in the present experiment is nonetheless striking. It is likely that the typographic enhancement and the realization that the word was glossed prompted longer processing of the first instance of the word. Also, given that the gloss clarified the word when first encountered, subsequent instances of the word may have prompted much less processing because it was no longer puzzling.

Moving on to the question of whether any of the online reading measures predicted vocabulary uptake, the findings corroborate previous studies in that total attention given to the target words was positively associated with test performance. And yet, post-test performance appeared the best under the picture-only condition, even though the total amount of time spent on the targets and their glosses was not greater in that condition than it was in the other two conditions. This disconnect between eye-tracking data and post-test scores suggests is that pictorial glosses can lend a mnemonic advantage, but not because they

engender protracted visual processing of either the gloss or the pseudowords. Note that the uptake of information from visual scenes is very rapid compared to reading – meaning can be extracted from complex visual scenes in as little as 100 ms (Biederman, Rabonowitz, Glass, & Stacy, 1974), that is, within a single fixation. This advantage may be particularly acute for second-language learners, for whom the meaning of the picture should be more transparent than the textual definition.

The question remains, though, why the combination of pictorial and textual representation of word meaning (in the multimodal gloss condition) was not also significantly more effective than the text-only gloss condition. After all, participants in the multimodal gloss condition also looked at the pictures, according to the eye-tracking data (although TFCs were lower and TRTs shorter than when only a picture was available). One possible explanation is that the presence of text (which might have been relatively hard for them to understand) may have interfered with the extraction of meaning from the picture. Another is that the interpretation of the picture without a textual clarification took slightly more effort (hence perhaps the slightly longer time spent on the pictures), and this generated stronger memories (as would be predicted under models such as Levels of Processing; e.g., Craik & Lockhart, 1972).

Conclusion and Limitations

Our study aimed to evaluate the effects of three different types of marginal glosses (text-only, picture-only, and text and picture) on L2 learners' uptake of new words from reading. The novelty of the study was to examine not only learning outcomes (through post-tests) but also online reading processes (through eye-tracking measures). We examined whether gloss type influences reading behaviours, and, if so, whether these reading behaviours (especially the amount of attention given to target words and to the information contained in glosses) help to predict learning outcomes.

Two post-reading tests were administered: a productive-knowledge test, where participants were prompted to recall the form of the target words, and a receptive-knowledge test, where they matched target words with their meaning. No difference emerged between scores on the first test under the three gloss conditions. The second test, however, revealed an advantage for glosses that included a picture, especially for the picture-only glosses.

In general, post-test performance was positively associated with the amount of attention participants gave to the target words and their glosses during reading, corroborating previous studies where eye-tracking data were predictive of post-reading test scores. The eye-tracking measures in the present study revealed only minimal differences in online processing of the target words and their glosses across the three gloss conditions, although one striking finding here was that text-only glosses tended to be skipped (i.e., ignored) far more often than glosses containing a picture. Another finding was that the target word reiterated in the gloss itself attracts very little attention, even if it is typographically enhanced (using bold face, in this experiment). This suggests that, when readers realize a word comes with a gloss and make the effort to consult it, they will promptly turn their attention to the clarification of the word rather than to the instance of the word itself. Interestingly, though, the length of their first fixation on the word in the gloss does appear to positively influence their retention of the word's form.

Altogether, the eye-tracking data provided little to explain why post-test performance tended to be the best after learners had read the text accompanied by picture-only glosses. This suggests that more facets of online processing play a part in establishing memories for words than meet the eye (-tracker). These might include contextual factors, although in the current case the text contexts at least were the same for each gloss condition. It is of course possible that the eye-tracking measures we opted to use here were insufficient to pick up relevant differences in participants' allocation of attentional resources. Deploying a larger arsenal of measures, including transitions (saccades) between areas of interest, might reveal additional differences.

Several other limitations to this study must be acknowledged. One concerns sample sizes, not only with regard to the number of participants but also with regard to the small number of target words. The fact that only two tests were used to measure learning outcomes is another limitation, because these may not have been sensitive enough to pick up differential learning gains at a subtle level of word-knowledge development. Finally, although the purpose of this study was not to measure the effects of glossing *per se* but rather to compare the effects of different gloss types, data from a control condition without any glossing could have been informative to evaluate the extent to which glossing as such influences reading behaviour. Despite these limitations, we hope this study will stimulate further investigations of text manipulations (such as glossing) intended to foster vocabulary uptake, and where online reading measures shed light on off-line learning outcomes.

References

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology, 40*, 23–31.

Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science, 35*(6), 499–517.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi:http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious Mixed Models*. Available from arXiv:1506.04967 (stat.ME). Retrieved from https://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4 (Version 1.1-8). Retrieved from http://CRAN.R-project.org/package=lme4

Biederman, I., Rabonowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*(3), 597-600.

Boers, F., Warren, P., Grimshaw, G., & Siyanova-Chanturia, A. (2017). On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning, 30*(7), 709–725. doi:10.1080/09588221.2017.1356335

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System, 66*, 113–129.

Brusnighan, S., & Folk, J. (2012). Combining contextual and morphemic cues is beneficial during incidental vocabulary acquisition: Semantic transparency in novel comound word processing. *Reading Research Quarterly, 47*(2), 172–190.

Chaffin, R., Morris, R., & Seely, R. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology Learning Memory and Cognition, 27*(1), 225–235.

Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research, 21*(3), 403-426.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*, 183–198.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671−684.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual Word Learning during Reading in a Second Language. *Studies in Second Language Acquisition*, 1-26. doi:10.1017/s0272263117000109

Eyelink Data Viewer Version 1.11.9000. (2007). [computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

Godfroid, A., Boers, F., & Housen, A. (2013). An Eye for Words: Gauging the Role of Attention in Incidental L2 Vocabulary Acquisition by Means of Eye-Tracking. *Studies in Second Language Acquisition, 35*(3), 483–517. doi:10.1017/S0272263113000119

Godfroid, A., & Schmidtke, J. (2013). What do eye-movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honour of Richard Schmidt* (pp. 183–205). Honolulu, HI: University of Hawai'i.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied,, 18*(2), 178–191.

Jones, L. C., & Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in French with multimedia annotations. *The Modern Language Journal, 86*(4), 546–561.

Khezrlou, S., Ellis, R., & Sadeghi, K. (2017). Effects of computer-assisted glosses on EFL learners' vocabulary acquisition and reading comprehension in three learning conditions. *System, 65*, 104-116. doi:10.1016/j.system.2017.01.009

Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, *46*, 56–79.

Kost, C. R., Foss, P., & Lenzini, J. J. (1999). Textual and pictorial glosses: Effectiveness of incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals, 32*, 89–113.

Lenth, R. V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software, 69*(1), 1–33. doi:10.18637/jss.v069.i01

Magnusson, A., Skaug, H., Nielsen, A., Berg, C., Kristensen, K., Maechler, M., . . . Brooks, M. (2017). glmmTMB (Version 0.1.1) Generalized Linear Mixed Models using Template Model Builder. Retrieved from https://github.com/glmmTMB

Mason, L., Pluchino, P., & Tornatora, M. C. (2016). Using eye-tracking technology as an indirect instruction tool to improve text and picture processing and learning. *British Journal of Educational Technology, 47*(6), 1083–1095. doi:10.1111/bjet.12271

Mayer, R. E. (2009). *Multimedia learning* (2nd. ed.). New York: Cambridge University Press.

Mohamed, A. A. (2017). Exposure Frequency in L2 Reading. *Studies in Second Language Acquisition*, 1-25. doi:10.1017/s0272263117000092

Murray, W. S. (1998). Parafoveal pragmatics. In G. Underwood (Ed.), *Eye Guidance in Reading and Scene Perception* (pp. 181–199). Amsterdam.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading. *Studies in Second Language Acquisition, 38*(1), 97–130.

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, 9*, 25–36.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2003). Modeling eye movements in reading: extensions of the E-Z Reader model. In J. Hyönã, R. Radach, & H. Deubel (Eds.), *The mind's eye: cognitive and applied aspects of oculomotor research* (pp. 361–390). Amsterdam: Elsevier.

Psychology Software Tools. (2012). E-Prime (Version 2.0). Retrieved from http://www.pstnet.com

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing [http://www.R-project.org/].

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention during reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*, 1457-1506.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of*

experimental psychology. Human perception and performance, *37*(2), 514–528. doi:10.1037/a0020990

Singmann, H., Bolker, B., & Westfall, J. (2015). afex: Analysis of Factorial Experiments (Version 0.13-145). Retrieved from https://cran.rproject.org/web/packages/afex/index.html

SR Research Experiment Builder Version 1.10.165. (2011). [computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 1*, 287–307.

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*, 312–339.

Winke, P. M. (2013). The Effects of Input Enhancement on Grammar Learning and Comprehension. *Studies in Second Language Acquisition, 35*(02), 323-352. doi:10.1017/s0272263112000903

Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of picture and annotation types. *CALICO Journal, 20*, 33–58.

Figure captions

Figure 1 Two-way interaction of Instance and Condition for First Fixation Durations on the pseudowords. (Means and standard errors. Values shown are those predicted by the mixed effects model, back-transformed from log values to millisecond values for transparency.)

Figure 2 Two-way interaction of Instance of the pseudoword and First Fixation Duration in predicting form recall score. (Values shown are those predicted by the mixed effects model. Shaded ribbons show standard errors.)

Tables

Table 1 Eye-tracking measures (FFD=mean First Fixation Duration, in milliseconds; Count=mean Fixation Count; Skip=skip rate in %; TRT=mean Total Reading Time, in milliseconds; TRT′=mean Total Reading Time, conditional upon the region not being skipped. The first four lines in each condition refer to instances of the pseudowords).

Gloss condition	Interest area	FFD	Count	Skip rate	TRT	TRT′
Text only	First in text	267	3.72	0.00	1022	1022
	In gloss	249	1.88	25.56	481	646
	Second in text	306	2.03	4.44	570	597
	Third in text	291	2.47	10.00	660	734
	Definition area	237	8.58	20.00	2244	2805
	Picture area	Not applicable				
Picture only	First in text	282	3.82	1.75	1061	1080
	In gloss	259	2.68	18.42	698	856
	Second in text	269	2.22	5.23	607	641
	Third in text	285	2.89	6.14	772	822
	Definition area	Not applicable				
	Picture area	264	2.02	17.54	570	692
Multimodal	First in text	285	3.35	1.85	953	970
	In gloss	227	2.44	12.96	598	687
	Second in text	266	2.24	4.63	594	623
	Third in text	249	2.29	5.56	594	629
	Definition area	220	8.36	11.11	2183	2456
	Picture area	244	1.29	24.07	335	442

Table 2 Post-reading word-form recall and meaning recognition by gloss condition (max = 1 in both cases).

Gloss condition	Form recall	Meaning recognition
Text only	0.08	0.21
Picture only	0.20	0.43
Multimodal	0.15	0.35

Appendices

Appendix A. Examples of text and glosses for the pseudoword *perchant*. In order, these are text-only, picture-only and multimodal gloss conditions

He was dressed in just light trousers, a T-shirt and a **perchant**, but he had a backpack containing a compass, bottled water, and some nut bars.
Luke battled through the difficult terrain for five hours.
"I was so happy that I was wearing a perchant," he said. He knew nights could get really cold in the bush and if he didn't find his way before dark, he would need it!
Luke knew by now that he shouldn't be looking for the panipline anymore but instead he should keep going in the same direction until he found a sign of human life, so he kept an eye on his compass to make sure he wasn't walking round in circles.

*a **perchant** is a jacket without sleeves*

He was dressed in just light trousers, a T-shirt and a **perchant**, but he had a backpack containing a compass, bottled water, and some nut bars.
Luke battled through the difficult terrain for five hours.
"I was so happy that I was wearing a perchant," he said. He knew nights could get really cold in the bush and if he didn't find his way before dark, he would need it!
Luke knew by now that he shouldn't be looking for the panipline anymore but instead he should keep going in the same direction until he found a sign of human life, so he kept an eye on his compass to make sure he wasn't walking round in circles.

*a **perchant***

He was dressed in just light trousers, a T-shirt and a **perchant**, but he had a backpack containing a compass, bottled water, and some nut bars.
Luke battled through the difficult terrain for five hours.
"I was so happy that I was wearing a perchant," he said. He knew nights could get really cold in the bush and if he didn't find his way before dark, he would need it!
Luke knew by now that he shouldn't be looking for the panipline anymore but instead he should keep going in the same direction until he found a sign of human life, so he kept an eye on his compass to make sure he wasn't walking round in circles.

*a **perchant** is a jacket without sleeves*

Appendix B. Example of the post-test used to measure recall of word form. See text for details.

4. Luke was wearing a short jacket without sleeves over his T-shirt. Can you remember the word that is used in the text?

Here are a sentence from the text and a picture that may help you remember:

"I was so happy that I was wearing a _____."
He knew nights could get really cold.

Appendix C. Example of the matching task used to test meaning retention. See text for details.

Which definition/picture below match the following word from the text?

perchant = *definition/picture #___*

1. short jacket without sleeves

2. small knife with a blade that folds back into the handle

3. painful local swelling that is caused by something repeatedly rubbing your skin

4. policeman of the lowest rank

5. wall built across a river that stops the water

6. footwear for long walks in the country

7. sticking plaster used to cover small wounds

8. bush with small yellow flowers

9. small sharp points on a plant

10. device used to find directions by means of a magnet

11. large farm building used for storing materials