8-1-2017

On the Benefits of Multimodal Annotations for Vocabulary Uptake from Reading

Frank Boers
Western University, fboers@uwo.ca

Paul Warren
Victoria University of Wellington

Gina Grimshaw
Victoria University of Wellington

Anna Siyanova-Chanturia
Victoria University of Wellington

Citation of this paper:
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.

On the Benefits of Multimodal Annotations for Vocabulary Uptake from Reading

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

Abstract

Several research articles published in the realm of Computer Assisted Language Learning (CALL) have reported evidence of the benefits of multimodal annotations, i.e., the provision of pictorial as well as verbal clarifications, for vocabulary uptake from reading. Almost invariably, these publications account for the observed benefits with reference to Paivio's *Dual Coding Theory*, suggesting it is the visual illustration of word meaning that enhances the quality of processing and hence makes new words more memorable. In this discussion article, we explore the possibility that it is not necessarily the multimodality *per se* that accounts for the reported benefits. Instead, we argue that the provision of multimodal annotations is one of several possible means of inviting more and/or longer attention to the annotations — with amounts of attention given to words being a significant predictor of their retention in memory. After reviewing the available research on the subject and questioning whether invoking Paivio's *Dual Coding Theory* is an optimal account for reported findings, we report an eye-tracking study the results of which are consistent with the alternative thesis that the advantage of multimodal glosses for word learning lies with the greater quantity of attention these glosses attract in comparison with single-mode glosses. We conclude with a call for further research on combinations and sequences of annotation types, regardless of multimodality, as ways of promoting vocabulary uptake from reading.

Introduction

A common means of assisting reading comprehension in a second language (L2) is the provision of annotations (or glosses) which clarify the meaning of unfamiliar words. Provision of these annotations is can also facilitate learners' uptake of the meaning of those new words (Hulstijn, Hollander, & Greidanus, 1996; for an early meta-analysis in the context of CALL, see Abraham, 2008). The way word meaning is clarified in annotations varies, and may involve first language (L1) translations, L2 synonyms or definitions, visual illustrations, or combinations of these elements. Several studies published in the realm of CALL in the past couple of decades have reported data in favour of using multimodal annotations (also known as multimedia annotations), i.e., annotations which provide textual and pictorial clarification of words, as a means of enhancing vocabulary uptake from reading (Akbulut, 2007; Al-Seghayer, 2001; Chun & Plass, 1996; Jones & Plass, 2002; Yeh & Wang, 2003; Yoshii, 2006; Yoshii & Flaitz, 2002). All of these included a comparison of uptake from text-only annotations (L1 translations or L2 synonyms/definitions) and uptake from annotations which also provide a visual illustration of the word's referent. With some exceptions (Acha, 2009), the latter condition was in general found to be the more beneficial of the two (see Chun, 2011, pp. 139–140; Mohsen & Balakumar, 2011; Xu, 2010, and Yun, 2011, for narrative or meta-analytic reviews indicating better post-test outcomes, overall, for reading conditions where words were annotated visually as well as textually).

Almost invariably, the authors of aforementioned publications invoke Allan Paivio's well established *Dual Coding Theory* (DCT) (Paivio, 1971; 1986; 1991) as a way of accounting for the attested benefits of multimodal annotations for vocabulary

uptake (Akbulut, 2007; Al-Seghayer, 2001; Chun & Plass, 1996; Jones & Plass, 2002; Mohsen & Balakumar, 2011; Xu, 2010; Yoshii & Flaitz, 2002; Yoshii, 2006; Yun, 2011). This is essentially an account which relies on the assumption there is a qualitative difference between the processing of text-only input and the processing of input that provides both textual and pictorial representations of word meaning. In this article, we ask (a) whether DCT is really an optimal frame of reference for this particular line of research on L2 vocabulary acquisition from annotated reading, and (b) whether an alternative explanation for the attested benefits of multi-modal annotations might be available. The latter explanation, we suggest, simply lies with the amount of attention that different annotation formats or annotation procedures attract.

In what follows, we briefly discuss Paivio's DCT and why we believe it does not apply optimally to the topic at hand. We then review empirical studies that have reported superior word retention after provision of multimodal annotations in comparison to text-only annotations, and argue that these did not consider the question of whether the amount of attention given to the annotations was comparable across treatment conditions. This is followed by a report of a new experimental study which used eye-tracking technology to address precisely that question.

Background

DCT (Paivio, 1986; 1991) is a theory of cognition which holds that there are two types of mental representations of concepts: verbal and nonverbal. Abstract concepts are typically coded only verbally, whereas concrete concepts are typically coded nonverbally as well as verbally, with mental imagery being the most common of nonverbal representations. A corollary of this is that the mental lexicon has two types of

lexical items: abstract-meaning items (e.g., *belief*) and concrete-meaning items (e.g., *apple*). The latter are concrete by virtue of their imageability — they can readily call up a mental picture of the thing, action or feature they refer to (Paivio, 2010).

DCT was inspired in part by the results of memory experiments which have revealed that series of concrete-meaning words tend to be easier to recall than series of abstract-meaning words (Paivio, 1965; Walker & Hulme, 1999). In addition, experiments have revealed the phenomenon of 'picture superiority': A series of pictures tends to be easier to recall than a series of words (Nelson, Reed, & Walling, 1976; Paivio & Csapo, 1977), which suggests that a nonverbal representation in memory offers an advantage for retrieval. From the combination of these two observations, it is reasonable to assume that, if concrete-meaning words are comparatively easy to retrieve from memory, this is thanks to their imageability, i.e., due to the ease with which they call up a mental image of their referents.

It needs to be borne in mind, however, that in the majority of these experiments the stimuli for recall tasks were familiar L1 words or pictures eliciting familiar L1 labels. So, when participants were asked to recall a series of pictures that had been shown to them, they already knew the words for the referents depicted therein. While the findings of these experiments support the key tenets of DCT, they do not as such reveal much about the role of pictures in the acquisition of *new* words, such as new words in a L2. Paivio and Lambert (1981) did demonstrate significant effects of imagery also in balanced bilinguals' recall of words, consistent with the predictions of DCT, but this again concerned already known words. There indeed seems to be no reason why DCT should not be extended to the balanced bilingual lexicon. For example, if the meaning of L1 *apple* is relatively easy to retrieve from memory by virtue of its

concreteness — it is easy to call up a mental image of the fruit — then so will the meaning of, say, the known L2 French counterpart *pomme*, because this has the same, highly imageable, referent as *apple*.

However, the line of research we examine in this article — the use of pictures in addition to verbal annotations in L2 reading — concerns the acquisition of *new* words, not recall of already familiar words. A crucial aspect of word learning is, of course, remembering what the word refers to, but word learning also requires remembering the phonological and orthographic form of the word (Nation, 2013). While a picture may serve as a direct means of elucidating word *meaning*, it is not obvious how looking at a picture as such should make the phonological and orthographic *form* of its corresponding word particularly memorable. In fact, some research suggests that pictures have the potential to distract from word form (Boers, Piquer Píriz, Stengers, & Eyckmans, 2009, Boers, Warren, He, & Deconinck, 2017; Carpenter & Olson, 2012). The use of pictures therefore seems more likely to benefit learners' retention of word meaning than their retention of word form. Paivio himself actually pointed this out as well: "Given that the image is retrieved, some items can be recalled simply by decoding the image. This would be the case especially of the more familiar second-language items, which are already represented and connected to referent images in long-term (semantic) memory." (Paivio & Desrochers, 1981, p. 788).

It is indeed almost exclusively tests of meaning recall or meaning recognition (where the newly encountered L2 word form is used as a cue) that have been used in the research on word annotations to date and that have produced evidence in favour of adding pictures. However, even with respect specifically to retention of the meaning of new words, it is not sure whether a DCT account works very well for the findings. This

is because the target words in the studies were generally concrete-meaning words referring to familiar objects, actions or features. In fact, in most cases when designers of pedagogic materials opt to elucidate the meaning of an L2 word by means of a picture, this option is available to them precisely because the word *has* a concrete meaning (e.g., Mohsen, 2016, 1225; Yoshii & Plaitz, 2002, 38). In cases where an adequate mental image is already available for a given referent (as in the case of *apple / pomme*, for instance), then it is not obvious (under a DCT account) why presenting learners with a picture of that referent should enhance its memorability. According to DCT, concrete concepts are easier to remember and recall precisely because they are also coded nonverbally, as a mental image. Once the equivalence is established between, for example, *apple* and its French counterpart *pomme*, then the latter can reasonably be expected to call up the familiar mental image of the fruit that would also be called up by the L1 word *apple*. It is not clear in such cases of easily 'imageable' words why the addition of a picture to a verbal clarification of word meaning should render the new L2 word more memorable than a verbal clarification alone. It might therefore be worth reviewing the available studies at the item level and examine whether the attested benefits of multimodal annotations might be most pronounced in the case of word meanings that are not straightforwardly imageable, and where the addition of a picture does potentially make a difference for its representation in memory. Support for this possibility comes from research in the realm of deliberate vocabulary learning, where a study by Farley, Ramonda and Liu (2012), for example, found that pictorial elucidation of word meaning aided recall of relatively abstract-meaning words but not recall of concrete-meaning words.

To be clear, we are not questioning the general usefulness of multimodal annotations accompanying texts. As demonstrated by Mayer and associates (Mayer, 2009; Mayer & Anderson, 1992; Mayer & Gallini, 1990), visuals can be very helpful to promote comprehension of scientific texts and of technical manuals, where verbal explanations may not suffice to help readers 'picture' the described concepts or processes. Still, even regarding its benefits for reading comprehension, the evidence in favour of multimodal input is not entirely systematic (see Chun, 2011, pp. 140–143, for a review). A study by Ariew and Erçetin (2004), for example, found no beneficial impact of the addition of still pictures for L2 readers' text comprehension. Moreover, they found a significant *negative* effect for the addition of video clips about the subject of the text, suggesting that such additional input carries the risk of reducing readers' attention to the actual reading text.

Returning to the use of pictures to elucidate *word* meaning, it is undeniable that visuals can usefully elucidate referents with which a given learner is not yet familiar, for example because they are culture-bound. Without the support of visual illustrations, it may be hard for learners to create an adequate mental image of such unfamiliar referents solely on the basis of a brief verbal description. As already mentioned, the use of visual symbols can also lend a certain degree of imageability to otherwise abstract-meaning words and make these more memorable that way. However, in the studies which examined the benefits of multimodal annotations for vocabulary uptake from L2 reading, the target words were generally concrete-meaning words referring to real-world referents which were presumably familiar to the learners. As argued above, it is far from clear why a pictorial presentation of such words should bestow on their meanings the mnemonic benefit of a 'dual' mental representation, if these meanings — being

imageable ones — already *have* a nonverbal representation in the mind. Instead, it would seem it is the forging of an accurate *verbal* representation of the to-be-learned word that requires attention.

Let it be clear as well that the above discussion is not at all meant as a critique of Paivio's DCT itself. We do not question DCT as a useful model of the mental representation of concepts/words, and its explanation for the advantage of concrete-meaning words in recall experiments. We are only raising the question of whether invoking DCT in studies which examine multimodality in L2 vocabulary earning — as many authors are wont to do — is well justified. With a few exceptions, such as Paivio and Desrochers' (1979, 1981) evaluation of variants of the *Keyword Mnemonic*, i.e., intentional vocabulary learning techniques where mental imagery is used to create associations between familiar and to-be-learned words, Paivio's publications seldom focus on L2 vocabulary acquisition and, to the best of our knowledge, make no claims about the benefits of multimodal annotations for vocabulary uptake from L2 reading. In fact, where Paivio advocates the use of associative imagery as a component of mnemonic techniques, this mostly concerns learner-generated mental images, *not* illustrated text: "The image-based systems take advantage of the learner's knowledge of the word by teaching him or her to generate images of situations that represent the meanings of the units of the new language and to use imagery as a private rehearsal and study technique, *without the use of a laboratory, projector, or even illustrated text*" (Paivio & Desrochers, 1981, p. 790; our emphasis). In sum, it seems to us that Paivio's work has been cited too 'liberally' in the CALL literature on multimodal annotations and L2 vocabulary learning.

Still, there is no denying that multimodal annotations have been found to be more beneficial for L2 word learning than text-only annotations in the majority of the empirical studies which include a comparison of these two reading conditions (Yun, 2011). The question then is whether other factors than those associated with DCT may account for this.

Literature Review

As already mentioned, much of the evidence in favour of multimodal annotations comes from studies conducted in the realm of CALL, where learners are asked to read L2 text on a computer screen. The multimodality of annotations has been operationalized in this body of research in two ways. One way is to signal to the reader that annotations are available and can be accessed by mouse-clicking on highlighted words, and this then gives them access to either only one annotation (e.g., a L1 translation) or two separate annotations (e.g., one a L1 translation and the other a picture). Chun and Plass (1996) and Jones and Plass (2000) are two studies which used this design and which report evidence that the provision of both a textual and a pictorial annotation tends to result in better retention of word meaning than the provision of just one annotation. While this result was interpreted as evidence of the benefits of multimodality, an alternative (or at least complementary) explanation is that the provision of two annotations invites two look-ups, whereas the provision of one annotation invites just one. As demonstrated in a similar study by Plass, Chun, Mayer and Leutner (1998), participants indeed tend to retain the meaning of target words better if they inspect two annotations about a word than if they inspect only one. Thus, any provision of two annotations (e.g., one an L1 translation and the other an L2 definition or example) could in theory stimulate two

look-ups instead of one, and could then in principle also result in better meaning retention, regardless of multimodality.

Akbulut (2007) is a more recent study where the participants' retention of the meaning of annotated words was also found to be better after reading a text in a condition where pictorial annotations were available in addition to textual ones than in a condition where only a textual annotation was available per target word. The author explains that computer software recorded how often participants accessed given annotations. Such data on individuals' look-up behaviour offer an excellent opportunity to try and statistically disentangle the effect of quantity of look-ups from the effect of multimodality *per se*. It is unfortunate, therefore, that this opportunity seems to have been missed — since the look-up data that were collected are *not* included in the article.

In the aforementioned studies, annotations were either verbal or pictorial, and to receive multimodal input, learners needed to visit two separate annotations. The second way in which multimodality is sometimes operationalized in this strand of work is to combine textual and pictorial input together in one gloss or annotation, which, according to Türk and Erçetin (2014), is the better option. An early but influential study of this kind is a pen-and-paper experiment by Kost, Fost and Lenzini (1999) where L2 learners read a short narrative text with marginal glosses for unfamiliar words. The glosses were either L1 translations, pictures, or a combination of translations and pictures. Vocabulary uptake was measured by means of diverse tests. Overall, the results showed an advantage for the gloss condition which combined translations and pictures. Later studies where pictures and verbal clarifications were combined in glosses were conducted in computer-aided reading contexts. For example, in a conceptual replication of Kost et al. (1999), Yoshii and Flaitz (2002) assigned students to one of

three annotation conditions: L2 definition only, picture only, or L2 definition plus picture. Again, the post-test results suggested an advantage of the multimodal glosses over the single-mode glosses (but see Boers et al., 2017, for a critique of the quality of the definition-only glosses in that study, where information necessary for the participants to perform optimally in the post-test was in some cases missing).

While research results have generally been favourable of the addition of pictures to annotations to promote vocabulary uptake from reading, there are some exceptions. In Acha (2009), for example, providing only L1 translations as annotations yielded significantly better post-test scores than providing multimodal annotations. Boers, et al. (2017) also compared learners' retention of word meaning after reading a text with either text-only or text-plus-picture annotations, and found no evidence to suggest superiority of the latter in any of the three experimental trials they conducted. It might be interesting to investigate to what extent inclusion of these studies in a new meta-analysis would reduce the (moderate) pooled effect size in favour of multimodal annotations as calculated by, for example, Yun (2011).

Still, the fact remains that the majority of published studies to date have claimed positive results for the use of glosses that combine textual and pictorial clarifications of word meaning. If we are correct in suggesting that invoking DCT does not offer a fully convincing explanation for the positive results attested in these studies, then we need to ask if an alternative (or at least complementary) explanation is available. What we propose is simple: multimodal glosses attract more attention from the reader than text-only glosses.

It is this possibility that we explore in the following section, where we report a small-scale experiment in which L2 readers' amount of attention to multimodal and single-mode glosses was gauged through the use of eye-tracking.

Experiment

Introduction

It is now commonly recognized that the amount of attention one gives to a new word while reading is one of the predictors of word learning. This is not only consistent with general theories in which attention is considered a prerequisite for learning (Schmidt, 2001), but it has also been empirically established by experiments where learners' eye movements during reading were recorded and where the number and duration of fixations on target words (as proxies of the amount of attention given to these words; Rayner, 1998) were found to be positively correlated with subsequent performance on word-meaning recall or word recognition tests (Godfroid, Boers, & Housen, 2013; Pellicer-Sánchez, 2016). What we illustrate by means of the experiment reported below is that marginal glosses that combine textual and pictorial information attract greater amounts of attention than marginal glosses that contain only the textual information.

We are by no means suggesting that amount of attention, as gauged by eye-movement data, is the sole (or even strongest) predictor of word uptake from reading. What mental operations one performs while attending to a word encountered in a text undoubtedly matters as well (e.g., Godfroid & Schmidtke, 2013). An influential proposal in this regard for estimating the role of mental engagement with new words is the *Involvement Load Hypothesis* (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001). According to this proposal, the likelihood of word learning from reading is greatest if

the learner (a) experiences a need to understand a novel word (for example, because comprehension of a text passage hinges on it), (b) then makes an effort to find information about the word (for example by looking it up in a dictionary), and (c) then evaluates whether the information found is satisfactory (e.g., whether it is compatible with the context where the word was met). In the below experiment, however, reading conditions were compared where all three components of the Involvement Load Hypothesis (i.e., 'need', 'search' and 'evaluation') were present to the same degree. The only difference between the reading texts was the nature of the glosses (i.e., text-only versus text + picture). One might nonetheless argue that the nature of the gloss has the potential to influence the learner's 'search' behaviour. The latter, at least, has been observed in experiments (e.g., Laufer & Hill, 2000) where L2 readers were found more inclined to inspect some types of hyperlinked annotations than others. Unlike those experiments, though, in the below experiment the participants were not presented with a choice between annotation types. Instead, word meanings were clarified in marginal glosses, accessible with minimal effort simply by casting one's eyes over them.

Participants, Materials, and Procedure

The experiment was conducted with the participation of adult international students with a high-intermediate level of proficiency in English, who were enrolled in an English proficiency programme at a university in New Zealand. Participation was entirely voluntary. After reading the call for volunteers, individual students contacted the research team, who then randomly assigned the students to one of the two reading conditions determined by the order in which they visited the lab to take part in the experiment. The experiment was run with one student at a time. The students were

informed the experiment was about reading in a second language, but its precise purpose was not revealed to them until data collection was completed.

The ESL learners were asked to read a text on computer screen while their eye movements were recorded by means of an EyeLink 1000 system (SR Research Ltd., Mississauga, Ontario, Canada). The text, an adaptation of a local news story, was approximately 900 words long and was distributed over eight screens. When participants finished reading a screen they pressed the space bar to move on to the next (but they could not return to the previous screen). Six pseudo-words were incorporated (each three times) in the story, all nouns with concrete meanings: *panipline* ('dam'), *perchant* ('jacket'), *dasters* ('nettles'), *hangles* ('blisters'), *bandilon* ('shed'), and *stavener* ('policeman'). The meaning of these pseudo-words was clarified in the right-hand margin of the text, opposite the line where the word first occurred in the text. So, readers did not need to click any hyperlinks to access the annotations. They just needed to temporarily shift their attention from the body of the text to the margin. After that first instance, each of the pseudo-words occurred twice more in the text, but without marginal gloss.

Two gloss conditions were created. In one version of the text, read by 15 participants, the glosses consisted of the target word (in bold) followed by a brief definition (e.g., *a **panipline** is a wall built across a river that stops the water*). In the second version, read by 18 participants, the glosses consisted of exactly the same textual information but additionally a picture of the referent (e.g., a dam) was provided immediately above it. A screen shot of each of the reading conditions is provided in the Appendix.

After the participants had finished reading the story, they were given a word-meaning recognition test. Each of the pseudo-words was presented on a separate screen, and the participants were asked to match the word with one of 11 referents to choose from, including the correct referent, the five referents of the other pseudo-words as lures, and five fillers.

Results

Fixation times and durations were analysed with regard to three particular areas of interest in the glosses: (a) the target word itself, (b) the definition, and (c) the picture. The eye-tracking measures deemed most relevant for the question addressed here, however, are the total number of fixations within an area of interest ('fixation counts') and the total amount of time spent on the area of interest over one or more fixations on that area ('total reading time'). Eye fixations were extracted using Eyelink Data Viewer Version 1.11.9000 (2007). Although eye-tracking technology is becoming more sophisticated and results more precise, there are still occasional issues when initial calibration is not perfect. To correct for this, the DataViewer's semi-automatic 'drift' correction algorithm was used. In any case, as said, of primary interest here is the readers' fixations in a distinct area—the marginal gloss—, and these can safely be taken as evidence that a reader was indeed looking at the gloss rather than something else.

Tables 1 and 2 show the average fixation counts and the average total reading times recorded under the two gloss conditions. Recall that these averages are based on 90 'observations' (15 participants x 6 glosses) in the text-only gloss condition and 108 'observations' (18 participants x 6 glosses) in the multimodal gloss condition.

<Table 1 about here>

<Table 2 about here>

The total number of fixations and the total reading times in the gloss area were entered as the dependent variables in linear mixed effects regression models (fitted with the lme4 package in R, Bates et al. 2015), with gloss condition as the fixed effect and participants and pseudo-words as random effects. For total fixation counts the effect of gloss condition approached conventional significance ($t = 1.95$, $p = 0.06$). A trend in the same direction was observed for total reading times, but also this fell short of significance ($t = 1.47$, $p = 0.15$). It is, of course, important to bear in mind that these computations are based on a relatively small data set, which inevitably yields conservative estimates of probability (or *p*-values).

If the multimodal glosses tended to attract more attention than the text-only glosses, this is not surprising, given that the addition of a picture is almost bound to make the former visually more salient (if only because the gloss occupies a larger area on the screen). In that sense, then, the nature of a marginal gloss may indeed exert an influence on readers' look-up behaviour—or, perhaps put more aptly here, their 'looking' behaviour. Put differently, the nature of the marginal gloss may arguably influence the 'search' element put forward in Laufer and Hulstijn's (2001) Involvement Load Hypothesis. It is pertinent in this regard that the average fixation counts and total reading times presented in Tables 1 and 2 are so markedly higher for the multimodal glosses due in part to the fact that the text-only glosses were *ignored* (i.e., not fixated at

all) as often as 20% of the time. The multimodal glosses were also ignored sometimes, but that happened 'only' 7.4% of the time.

If we exclude the 'non-fixations' from the comparison, the average total reading time nevertheless remains shorter for the text-only than the multimodal glosses (2,882 milliseconds and 3,275 milliseconds, respectively). So, also when they *are* looked at, text-only glosses appear to attract slightly less attention than do multimodal glosses. Again, this is not particularly surprising as the latter contain an extra component that invites attention—the picture. When it was looked at, the average total reading time on the picture alone was 467 milliseconds. This corresponds roughly to the size of the difference in total reading time between the two conditions.

Turning now to the post-reading test, where the participants were asked to match the words with their meanings, the results revealed an average success rate of 21% under the text-only gloss condition and of 35% under the multimodal gloss condition. The direction of the difference parallels that of the fixation data, as expected. A chi-square test of the distributions of correct and wrong word-meaning matches showed that the proportion of correct responses was significantly greater in the multimodal condition ($\chi^2 = 4.74$, $p < .05$). The difference between the two conditions thus looks even more pronounced in the case of the vocabulary test results than it did for the eye-tracking data. This, one might argue, indicates that amount of attention alone cannot suffice as an explanation for the superior test results under the multimodal gloss conditions. While we do not wish to exclude that possibility, it is worth pointing out that the large difference in proportion of correct post-test responses may to some extent be an artefact of the test format used. If one remembers the meaning of just one or two of the six pseudo-words, then the chances of lucky guesses when it comes to the remaining

pseudo-words are slim. If one remembers the meaning of three or four of the six pseudo-words, then the chances of successfully guessing the meaning of the few remaining ones increase considerably. In other words, a relatively small advantage for the multimodal gloss condition may have been inflated due to the test format used here.

Discussion, Conclusion, and Perspectives

We have argued that the benefits of text-plus-picture annotations for vocabulary uptake from reading that have been reported in several publications are not necessarily a reflection of the pictures' promotion of the words' concreteness or imageability, because it is highly likely that the referents of most of the target words in these published studies *are* already represented in memory as concrete and imageable. If so, the recurring references in this strand of research to Paivio's *Dual Coding Theory* to account for the proclaimed superiority of multimodal over text-only annotations may not be totally justified. The alternative—or at least complementary—account we have proposed is that the attested benefits lie with the amount of attention that multimodal clarifications tend to attract. In the case of separate annotations with either textual or pictorial clarifications, learners are likely to inspect word meaning more than once, resulting in stronger memory traces. This, rather than a DCT account, could explain the results of studies such as Chun and Plass (1996), Jones and Plass (2000) and Akbulut (2007). In cases where textual and pictorial clarifications are combined in one complex gloss, learners are more likely to attend to the gloss and to inspect it for longer in comparison to a simpler gloss. This, then, may be an alternative explanation for the results of studies such as Kost et al. (1999) and Yoshii and Flaitz (2002).

The results of a CALL-situated study consistent with the tenet that longer attention to glosses is likely to be associated with better retention is Al-Seghayer (2001), where three annotation types were compared: a verbal clarification only, a verbal clarification accompanied by a still picture, and the same verbal clarification accompanied by a video clip. The latter condition appeared most beneficial. This is consistent with our thesis, because watching a video clip is expected to take longer than glancing at a still picture. Note that in this study, the presence of multimodality as such cannot explain the different results obtained for the still-picture and the video annotations, since both were multimodal. On the other hand, a study by Yeh and Wang (2003) has demonstrated that more is not always better, and thus appears to contradict our suggestion. Also in that study, three annotation types were compared: a verbal clarification, a verbal clarification accompanied by a still picture, and an annotation where in addition to the verbal clarification and the still picture there was an audio recording with a demonstration of the word's pronunciation and spelling. The latter was found the least beneficial according to post-reading tests. The tests focused on word meaning, however, and did not require knowledge of the pronunciation of the target words. It is therefore not so surprising that the audio component of the annotations, which directed attention to phonological form rather than meaning, did not assist learners' test performance. Rather, the focus on word form may even have come at the expense of learners' engagement with word meaning (Barcroft, 2015), especially when time pressure is imposed upon the learners (Yeh & Wang, 2003, p. 240). This certainly adds a nuance to the suggestion we are making in this article: It is not necessarily total time spent on an annotation *as a whole* that matters, but rather the time spent on those

components of the annotation that will serve a particular learning purpose (and thus performance on a particular test format) best.

One of the conclusions reached by Schmitt (2008), in a review of research on L2 vocabulary acquisition, is that the amount of attention (or what he terms 'engagement') invested in new words is a strong predictor of learning. There is no denying, of course, that the nature of the mental operations performed while attending to words also matters (Eyckmans, Boers, & Lindstromberg, 2016; Barcroft, 2015; Craik & Tulving, 1975), and so quantity of attention (as gauged by means of eye-tracking) is far from the sole predictor of learning (e.g., Godfroid & Schmidtke, 2013). In the same vein, there is no denying that visuals *can* make a difference in learners' vocabulary retention (Boers, et al., 2009; Farley et al., 2012), but perhaps especially so where the mental images do not already spring to mind automatically (as in the case of familiar, concrete referents). That is where we feel DCT is definitely helpful as an explanatory framework.

Again, it is important to emphasize that the intention of this article is not to discard the usefulness of pictorial annotations in addition to textual ones, even in cases where the latter alone suffice to clarify word meaning. If the presence of visuals promotes learners' attention to the meaning of target words and if this in turn promotes uptake—as the results of our small-scale experiment suggest—, then providing these visuals must be welcome. However, there are several ways of designing multiple annotations or 'rich' glosses which could in theory serve this purpose, regardless of multimodality. Might a reading condition where words are annotated twice in a textual mode not generate similar results as a condition where words are annotated once in a textual mode and once in a pictorial mode? To the best of our knowledge, this possibility has not yet been empirically explored. From a pedagogical perspective, it

certainly seems worth investigating how a sequence of annotation (or gloss) forms can be employed so as to extend learners' mental engagement with target words (and to foster knowledge of various facets of these words). For example, it is not hard to envisage a computer-aided reading condition where a target word occurs several times in a text, each time with a link to an annotation of a different kind: On the first occurrence the annotation might be a verbal clarification; on the second it might be a picture, on the third it might be a recording of the word's pronunciation, and on the fourth the annotation might include a retrieval challenge (such as choosing the correct meaning of the word in a multiple-choice task; Nagata, 1999; Watanabe, 1997). Research designs in which the effectiveness of such different combinations and different sequences of annotations are assessed could at the same time shed more light on the added value of pictorial representations to L2 learners' vocabulary development.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback on an earlier version of this paper. We would also like to thank Murielle Demecheleer for her assistance with the design of the reading materials used in the experiment, and Ross van de Wetering and Angus Chapman for their assistance with the collection of eye-tracking data. The study reported in this article is part of a larger project that received financial support from University Research Fund of Victoria University of Wellington, New Zealand (URF Grant number 204092).

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 21*, 199–226.

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*, 499–517.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology, 5*, 202–232.

Ariew, R., & Erçetin, G. (2004). Exploring the potential of hypermedia annotations for second language reading. *Computer Assisted Language Learning, 17*, 237–259.

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. Amsterdam, Netherlands: John Benjamins.

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

Boers, F., Piquer Píriz, A.M., Stengers, H., & Eyckmans, J. (2009). Does pictorial elucidation foster recollection of figurative idioms? *Language Teaching Research, 13*, 367–388.

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

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 92–101.

Chun, D. M. (2011). CALL technologies for L2 reading post Web 2.0. In N. Arnold & L. Ducate (Eds.), *Present and Future promises of CALL: From theory and research to new directions in language teaching* (pp. 343–354). London: Routledge.

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., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Eyckmans, J., Boers, F., & Lindstromberg, S. (2016). The impact of imposing processing strategies on L2 learners' deliberate study of lexical phrases. *System, 56*, 127–139.

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

Farley, A. P., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research, 16*, 449–466.

Godfroid, A., & Schmidke, 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 honor of Richard Schmidt* (pp. 183-

205). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition, 35*, 483–517.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*, 539–558.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80*, 327−339.

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

Kost, C. R., Fost, 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.

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 and Technology, 3*, 58–76.

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

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

Mayer, R. E., & Anderson, R. W. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology, 84*, 444−452.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82*, 715−726.

Mohsen, M. A. (2016). Effects of help options in a multimedia listening environment on L2 vocabulary acquisition. *Computer Assisted Language Learning, 29,* 1220–1237.

Mohsen, M. A., & Balakumar, M. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition in CALL literature. *ReCALL, 23,* 135–159.

Nation, I.S.P. (2013). *Learning vocabulary in another language (2nd ed.)*. Cambridge: Cambridge University Press.

Nelson, D. L., Reed, V. S., & Walling, J. R. (1976). Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning & Memory, 2,* 523–528.

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired associate learning. *Journal of Verbal Learning & Verbal Behaviour, 4,* 32–38.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston. Reprinted as Paivio. A. (1979). *Imagery and verbal processes*. Hillsdale, N.J.: Erlbaum.

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

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*, 255–287.

Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon, 5*, 205–230.

Paivio, A., & Csapo, K. (1977). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5*, 176−206.

Paivio, A., & Desrochers, A. (1979). Effects of an imagery mnemonic on second language recall and comprehension. *Canadian Journal of Psychology, 33*, 17–28.

Paivio, A., & Desrochers, A. (1980). Mnemonic techniques in second-language learning. *Journal of Educational Psychology, 73*, 780–796.

Paivio, A., & Lambert, W. (1981). Dual coding and bilingual memory. *Journal of Verbal Learning & Verbal Behaviour, 20*, 532–539.

Pellicer- Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition, 38*, 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.

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

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge: Cambridge University Press.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329–363.

Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning, 27*, 1–25.

Walker, I., & Hulme C. (1999). Concrete words are easier to recall than abstract words: evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1256–71.

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.

Xu, J. (2010). Using multimedia vocabulary annotations in L2 reading and listening activities. *CALICO Journal, 27*, 311–327.

Yeh, Y., & Wang, C-W. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal, 21*, 131–144.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*, 85–101.

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

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning, 24*, 39–58.

Table 1: Average number of fixations in the gloss area

	Gloss condition	
	Text only	Text plus picture
Fixations on the target word	1.88	2.88
Fixations on the definition	6.70	7.11
Fixations on the picture	--	1.69
Total fixations within the gloss	**8.58**	**11.68**

Table 2: Average total reading times (in milliseconds) in the gloss area

	Gloss condition	
	Text only	Text plus picture
Total reading time on the target word	481	653
Total reading time on the definition	1,764	1,759
Total reading time on the picture	--	379
Total reading time within the gloss	**2,245**	**2,791**