

UNPACKING THE COMPARISON OF L1 AND L2 GLOSSES IN VOCABULARY
LEARNING FROM READING

By

Yingzhao Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2023

ABSTRACT

The appropriate amount of first language (L1) and second language (L2) to use in L2 learning has been constantly debated (e.g., Cummins, 2007; Hall & Cook, 2012). This study situates the debate of L1 and L2 use in the context of vocabulary learning from reading. By examining the potential moderating factors on the comparison of L1 and L2 glosses (i.e., short word definitions provided during reading), the study aims to provide a nuanced picture of how L1 and L2 input affects vocabulary learning in various circumstances. Investigating L1 and L2 glosses in the context of vocabulary learning also allows the study to contribute to the theories of bilingual lexicon (e.g., Kroll & Stuart, 1994; Jiang, 2000), i.e., the development of the bilingual lexicon as a function of input language.

One hundred and eighteen L2 learners of English completed the study. Participants first read part of a graded reader, where 24 target words were embedded. Glosses for the target words were inserted through hyperlinks: participants could click the target words to access their glosses, written either in the participants' L1 or L2. Participants' time spent on reading each gloss was tracked. After reading, participants went through unannounced vocabulary posttests that measured receptive and productive meaning knowledge, and lexical retrieval fluency of the target words. Participants also filled in an exit questionnaire that aimed to further probe their reading and gloss access behaviors. Participants' gloss reading time, their vocabulary size, and the target words' frequency of occurrence (FoO) were analyzed as moderating variables on the comparison of L1 and L2 glosses.

Findings revealed that the comparative effects of L1 and L2 glosses were primarily moderated by participants' gloss reading time and target word FoO, suggesting that the initial depth of processing and subsequent memory reactivation were the keys to successful vocabulary

learning. Results have pedagogical implications for how to choose the language for glosses, theoretical implications on the bilingual lexicon development, and methodological implications of using hyperlinks to track behaviors during learning.

ACKNOWLEDGEMENTS

I am lucky to have been surrounded by kind and supportive people throughout my PhD journey. In particular, I would like to express my gratitude to the following individuals, without whom this endeavor would not have been possible.

First, I am deeply indebted to my advisor, Dr. Shawn Loewen. Shawn has always supported my research endeavors, encouraging me to probe deeper and take on new adventures. He cheers for my success and never hesitates to offer emotional support during challenging times. He shares personal stories to help me navigate different aspects of academia. He is open-minded, humorous, and easy going. There are always smiles and laughs when working with him. I am also grateful to Dr. Aline Godfroid, for her guidance in my interdisciplinary research pursuit. Many thanks to other members of my committee: Drs. Meagan Driver and Sandra Deshors, for their constructive feedback on my dissertation and other research projects.

I hope to express my appreciation to three individuals who are not on my committee: Drs. Nan Jiang, Paula Winke, and Patti Spinner. Nan has inspired my research. I gain new insights into bilingualism research from every interaction with him. Paula is always willing to set aside time in her busy schedule to talk to me and offer me career advice. Patti is my mentor in teaching. She is encouraging and kind, giving me much freedom to explore new things.

I am lucky to have done my PhD in the Second Language Studies program. I would like to thank students and faculty members in the program for a supportive working environment.

This dissertation project would not have been possible without the financial support from the Duolingo Dissertation Grant, the National Federation of Modern Language Teachers Associations (NFMLTA) Dissertation Support Grant, the International Research Foundation for English Language Education (TIRF) Doctoral Dissertation Grant, the Mango Languages

Dissertation Award, and the Dissertation Completion Fellowship from the College of Arts and Letters and the Second Language Studies program. I would also like to acknowledge the time and effort of my participants.

Last but not least, words cannot express my gratitude and love to my family. I would not have survived all the challenges and made it this far without my family.

TABLE OF CONTENTS

INTRODUCTION	1
CHAPTER 1: LITERATURE REVIEW	3
CHAPTER: 2 METHOD	37
CHAPTER 3: RESULTS	60
CHAPTER 4: DISCUSSION AND CONCLUSION	100
REFERENCES	117
APPENDIX A: TASK INSTRUCTIONS.....	131
APPENDIX B: VOCABULARY PROFILE OF THE READING MATERIAL.....	134
APPENDIX C: TARGET WORD CHARACTERISTICS.....	135
APPENDIX D: GLOSSES	138
APPENDIX E: EXIT QUESTIONNAIRE	141
APPENDIX F: LANGUAGE BACKGROUND QUESTIONNAIRE.....	143
APPENDIX G: STIMULI FOR THE SELF-PACED READING TEST.....	150
APPENDIX H: MIXED-EFFECTS MODELLING	153

INTRODUCTION

Whether to use second language (L2) alone or both the first language (L1) and the L2 in L2 learning has long been a contentious issue. Theorists in instructed L2 learning have argued for a bilingual teaching approach while language policy makers seem to encourage maximal L2 use. Empirical findings from classroom studies indicated that L1 use is beneficial for vocabulary learning (e.g., Tian & Macaro, 2012; Zhao & Macaro, 2016) but may not be superior to exclusive L2 use for the learning of other aspects (Brown, 2021; Brown & Lally, 2019). In contrast to advocates of bilingual teaching in instructed L2 learning, three bilingual lexicon models, i.e., the Revised Hierarchical Model (the RHM; Kroll & Stewart, 1994), Jiang (2000), and the Revised Hierarchical Model-Repetition Elaboration Retrieval (RHM-RER; Rice & Tokowicz, 2020), predicted that L1 input is less conducive to developing high-quality lexical representations that would allow words to be retrieved fluently. This prediction has been supported by a number of psycholinguistics studies (e.g., Comesaña et al., 2009; Elgort & Piasecki, 2014; Jeong et al., 2010).

Given the contradiction in bilingual lexicon theories and in empirical findings in instructed L2 learning, this study set out to compare L1 and L2 use in the context of glossing (i.e., providing a short definition for a word) in vocabulary learning from reading. Current research on the comparison of L1 and L2 glosses not only has yielded inconsistent findings (H. S. Kim et al., 2020; Zhang & Ma, 2021) but is also limited in number (H. S. Kim et al., 2020) and scope, often overlooking important variables related to the learning condition and learner characteristics. In this study, I examined three variables that may potentially moderate the comparative effectiveness of L1 and L2 glosses, namely L2 vocabulary size, engagement with glosses, and target words' frequency of occurrence (FoO). The study extends previous gloss

language studies by including posttests that measured not only learners' ability to recognize and recall word meanings but also their fluency of word retrieval in reading.

The current study represents one of the first attempts to unpack factors that influence the effects of gloss language in vocabulary learning. Theoretically, the study contributes to the debate about how the bilingual lexicon develops as a function of input language. Pedagogically, findings shed light on how to best utilize glosses based on learner and target word characteristics.

CHAPTER 1: LITERATURE REVIEW

Use of First and Second Language in Second Language Learning

Second language (L2) instructors and learners have the liberty to choose the amount of first language (L1) and L2 to use in the learning of the L2. For example, instructors may explain L2 grammatical rules in learners' L1 and learners may look up the meaning of an L2 word in a monolingual (i.e., L2 only) dictionary. Whether to use L1 and how much L1 to use in L2 learning have been debated for decades. In this section, I first review classroom research on language of instruction. I then discuss the use of L1 and L2 input on the development of the bilingual lexicon. I focus on three bilingual lexicon models, namely the Revised Hierarchical Model (the RHM; Kroll & Stewart, 1994), Jiang's (2000) model, and the Revised Hierarchical Model-Repetition Elaboration Retrieval model (RHM-RER; Rice & Tokowicz, 2020).

Language of Instruction in L2 Classrooms

The two sides of the debate about language of instruction are monolingual and bilingual teaching. Monolingual teaching refers to using the L2 exclusively and bilingual teaching involves the use of both learners' L1 and L2. Several language teaching approaches have addressed the amount of L1 and L2 use in L2 classrooms. The grammar-translation approach, which emphasizes the learning of grammatical rules and the ability to translate written text, explicitly encourages language teachers to give instruction in learners' L1 (Celce-Murcia, 2014). Many have advocated translanguaging practices in language classrooms, which see learners' L1 and L2 as part of their linguistic repertoire and encourage learners to blend their multiple languages to achieve communicative goals (e.g., Canagarajah, 2011; Creese & Blackledge, 2010; De Costa et al., 2017). In contrast, monolingual teaching seems to stem from the view that L2 learning is similar to children acquiring their L1 and takes place when learners are immersed in

an L2-only environment (Cummins, 2007; De la Campa & Nassaji, 2009). For example, the use of L1 is explicitly denounced in the direct method, which focuses on meaning instead of grammatical rules and became popular as a response against the grammar-translation approach. Although communicative language teaching, an approach also with a focus on meaning, does not ban L1 use entirely, it argues for minimum L1 in L2 classrooms (V. Cook, 2001; Cummins, 2007). Theories aside, when it comes to real-world language teaching, language policy makers seem to favor the monolingual approach: the American Council on the Teaching of Foreign Language (ACTFL), in their guiding principles, recommends that teachers should use the L2 at least 90% of the time (n.d.); both the Japanese and Korean governments promote teaching English in English (see Kubota, 2018; Macaro & Lee, 2013).

The different opinions on L1 use in L2 learning reflect several beliefs. The first one pertains to the goal of language learning. In monolingual teaching, L2 learners' ultimate goal is to use the L2 to communicate with native speakers of the language in a monolingual environment while bilingual teaching aims for learners to use the L2 as a *lingua franca* in a multilingual environment (Hall & Cook, 2012). The second belief is on the relationship between the L1 and the L2. Monolingual teaching largely holds that the L1 and the L2 are separated, despite the now widely accepted view that the two languages are interdependent (V. Cook, 2001; Cummins, 2007; Hall & Cook, 2012) and that cognitive, academic, and literary skills in the L1 can be transferred to the L2 (Cummins, 2001). Related to the L1-L2 relationship is the third belief on the role of L1 in L2 learning. Because the L2 is seen as independent of the L1, monolingual teaching treats the L1 as an interference that must be eliminated from L2 learning. Proponents of monolingual teaching argue that only when the L1 is avoided can L2 exposure be maximized, which is crucial especially in a foreign language learning context, where learners do

not have much contact with the L2 beyond the classroom (e.g., Carless, 2007; Chambers, 1991; Chaudron, 1988; Turnbull, 2001). L1 use is also seen as a means of compensating inadequate L2 proficiency (Sato & Angulo, 2020). On the other hand, in bilingual teaching, the L1 and the L2 are interconnected and the L1 is viewed as a resource L2 learners can draw on. On a cognitive level, the use of L1 is believed to reduce processing loads on learners and thus facilitate learning (e.g., Hall & Cook, 2012; Scott & Fuente, 2008); L1 can also be used as a tool to mediate the process of problem solving (e.g., Antón & DiCamilla, 1998; Moore, 2013; Sato & Angulo, 2020; Watanabe, 2020). On a social level, L1 use helps preserve L2 learners' linguistic and cultural identities (G. Cook, 2010) and maintain linguistic diversity by resisting the dominance of English as well as English native speakers (Phillipson, 1992).

In terms of empirical research on the use of L1 and L2 in L2 classrooms, most studies are observational, documenting the amount of and reasons for L1 use. These studies have found that, theoretical debate aside, teachers in actual classrooms are likely to use learners' L1 in one way or another (e.g., Bruen & Kelly, 2017; De la Campa & Nassaji, 2009; Duff & Polio, 1990; Liu et al., 2004; Macaro, 2001; Polio & Duff, 1994; Tognini & Oliver, 2012). Teachers use the L1 for various reasons. Polio and Duff (1994) observed six foreign language classrooms in the US and identified eight functions of L1 use. Examples of these functions included pedagogy (e.g., grammar instruction, explaining difficult words, and ensuring comprehension), classroom management, and connecting with students (e.g., making jokes and expressing empathy) (see also Ma, 2019). De la Campa and Nassaji (2009) identified 14 uses of the L1 in two German-as-a-foreign-language classes. Out of these uses, translating L2 utterances in class was the most frequent one, followed by building rapport and making the learning atmosphere more comfortable. Nakatsukasa and Loewen (2015) examined the relationship between the amount of

L1 use and the linguistic areas being taught. They found that in teaching grammar and semantics, teachers used a similar amount of L1 and L2 while in vocabulary teaching, the majority of teacher discourse (60%) was in the L2. L1 is also common in peer interactions. Several studies examining spoken peer interaction have found that learners usually used the L1 for (1) language issues, such as finding the right word and using grammar correctly; (2) metacognitive purposes, such as goal setting and planning; and (3) social purposes, such as a casual discussion of unrelated topics (e.g., Gánem-Gutiérrez & Roehr, 2011; Storch & Aldosari, 2010; Swain & Lapkin, 2000; Tian & Jiang, 2021; Vraciu & Pladevall-Ballester, 2022; Xu & Fan, 2021). Yu and Lee (2014) looked into L1 use in peer interaction in written format. Learners in the study gave peer reviews in an L2 writing task. Results revealed that L1 was mostly used to give comments on the content of the essays while L2 feedback was mostly directed to form (e.g., vocabulary use and grammar). L2 proficiency has been shown to affect the amount of L1 use among peers, with lower-proficiency learners tending to use more L1 than their high-proficiency peers (e.g., DiCamilla & Antón, 2012; Xu & Fan, 2021; Yu & Lee, 2014).

A number of studies went beyond observation and examined learners' attitudes towards the inclusion of L1 in L2 learning. Most of these studies indicated that the use of L1 was welcomed by the majority of learners (e.g., Brevik & Rindal, 2020; Brooks-Lewis, 2009; J. H. Lee & Lo, 2017; Macaro et al., 2020; Tian & Hennebry, 2016; see Shin et al., 2020 for a review). Participants in these studies perceived L1 as facilitative in improving comprehension (e.g., Brooks-Lewis, 2009), reducing anxiety (e.g., Tian & Hennebry, 2016), and solving language problems (e.g., Macaro et al., 2020). Studies have also shown that learners, while agreeing that L1 should be included in L2 classrooms, also asked for more L2 use where possible. Macaro et al. (2020), for example, listed several areas where learners preferred L2 use over L1, including

giving instructions, explaining new words, asking and answering questions. Primary school learners in Nilsson (2020) expressed preference for predominant use of the L2, despite concerns over and actual experiences of difficulties in understanding class content in the L2. Like the amount of L1 use, attitudes towards inclusion of L1 were shown to be moderated by learners' L2 proficiency. In general, advanced learners preferred more L2 use than lower-proficiency ones (J. H. Lee & Lo, 2017; Tian & Hennebry, 2016). Macaro and Lee (2013) found that age also affected learners' perception of L1 use, with adult Korean learners being more likely to accept word definitions given in the L2 than young learners. The authors hypothesized that the age effect may be related to adult learners' higher L2 proficiency.

Fewer studies on language of instruction have directly examined how the use of L1 and L2 affected language development. Most of these studies focused on word learning (e.g., J. H. Lee & Levine, 2020; J. H. Lee & Macaro, 2013; Tian & Macaro, 2012; Zhao & Macaro, 2016). These studies unanimously revealed an advantage of incorporating L1 over using L2 only in classroom vocabulary learning, whether it was vocabulary learning through reading (J. H. Lee & Macaro, 2013; Zhao & Macaro, 2016) or listening (J. H. Lee & Levine, 2020; Tian & Macaro, 2012). H. Lee and Lee's (2022)'s meta-analysis on L2 vocabulary learning through teacher explanation showed that learning through L2 input yielded fewer gains than L1 input both in the short term and the long term. Zhao and Macaro (2016) suggested that learning L2 vocabulary through the L1 made retrieval of word meanings more straightforward, leading to the advantage of bilingual teaching. Qualitative analysis of their data also indicated that some students misunderstood or had difficulty understanding word meanings written in the L2, which was another possible reason for the disadvantage of learning vocabulary solely through the L2. J. H. Lee and Macaro (2013) found that the comparison of monolingual and bilingual teaching was

moderated by age in that young learners benefitted more from L1 use than adult learners, echoing findings in Macaro and Lee (2013) that young learners seemed to favored L1 use more than adult learners. The advantage of bilingual teaching over monolingual teaching was not altered by L2 proficiency (H. Lee & Lee, 2022.; Tian & Macaro, 2012) nor the concreteness of target vocabulary items (Zhao & Macaro, 2016). Two longitudinal classroom studies, Brown (2021) and Brown and Lally (2019), investigated monolingual and bilingual teaching in other areas of language. Results from these two studies were mixed regarding the comparative effectiveness of the two teaching approaches. No significant difference was found between the monolingual and bilingual teaching classes after 15 weeks of instruction in areas of writing and speaking (Brown & Lally, 2019). Similarly, beginner French L2 learners in the monolingual and bilingual teaching classrooms made similar progress in listening, reading, writing, and vocabulary after a 10-week course (Brown, 2021). Beginner Arabic L2 learners in the bilingual teaching class, however, made significantly more progress in vocabulary than those in the monolingual teaching class (Brown, 2021).

The abovementioned empirical studies suggest that the amount of, the attitudes towards, and the effects on learning of L1 and L2 use in L2 classrooms can be moderated by a number of factors, including L2 proficiency, age, and the target linguistic areas. This highlights the need for researchers to switch focus from the ‘whether’, that is, whether L1 should be used at all, to the ‘when’ and ‘how’, that is, choosing the language of instruction based on the target linguistic structures, learners’ individual characteristics, and other factors. With technology, it is now possible to adapt language instruction in real time based on learner performance and progress, making it more important to investigate the ‘when’ and ‘how’ of L1 and L2 use. The current study’s investigation on factors that affect the comparison of L1 and L2 glosses in a digital

learning environment fits into the broader trend of adaptive language learning. Examining the moderating variables on the effects of gloss language recognizes the roles of learners' own language, seeing a learners' multiple languages, i.e., L1s and L2s, as a repertoire of linguistic resources that can be deployed flexibly, rather than separate entities (Canagarajah, 2013; Creese & Blackledge, 2010, 2015; De Costa et al., 2017).

Input Language in Bilingual Lexicon Development

Research in the previous section on language of instruction was mostly classroom studies and came from the perspective of instructed second language acquisition. In this section, I review L1 and L2 use from a psycholinguistic perspective. Specifically, I focus on bilingual lexicon development and the effect of input language on it.

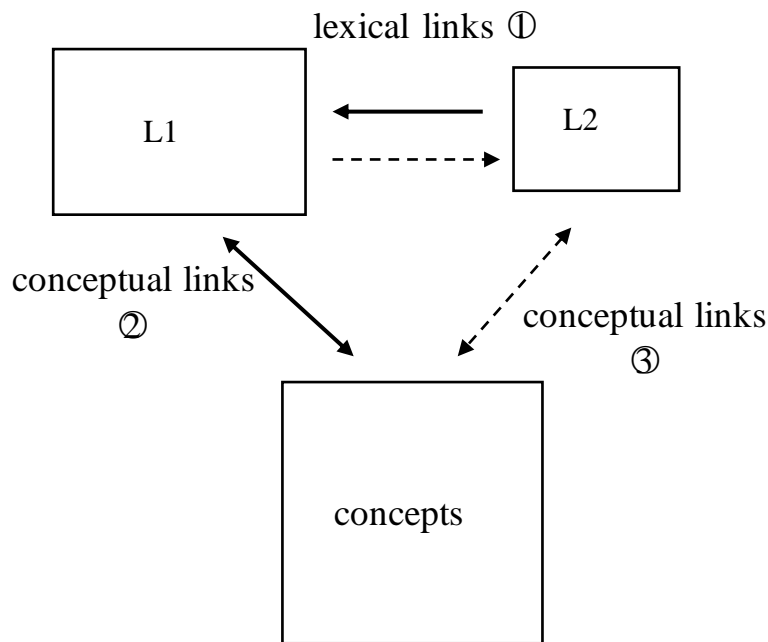
There are two key theoretical issues that most bilingual lexicon models attend to (see Dijkstra & van Heuven, 2002). The first one concerns the structure of the bilingual lexicon, that is, whether the L1 and L2 mental lexicons are separated or integrated. The second one discusses whether bilingual processing is selective or not, that is, whether words from only one language (i.e., selective) or both languages (i.e., nonselective) are activated. Connectionist models, e.g., the Bilingual Interactive Activation (the BIA; Dijkstra et al., 1998) and the BIA+ (Dijkstra & van Heuven, 2002) models, hypothesized an integrated lexicon and nonselective activation of L1 and L2 words. The Revised Hierarchical Model (the RHM; Kroll & Stewart, 1994) assumed separate lexicons for the L1 and the L2, but at the same time acknowledged nonselective lexical access (Kroll, Bobb, & Wodniecka, 2006; Kroll, van Hell, et al., 2010; cf. Brysbaert & Duyck, 2010). Jiang's (2000) model and the Revised Hierarchical Model-Repetition Elaboration Retrieval model (RHM-RER; Rice & Tokowicz, 2020) were founded on the RHM and posited similar views on the structure of L1 and L2 lexicons and bilingual processing. The RHM, Jiang (2000),

and the RHM-RER were chosen as theoretical support for the current study mainly for two reasons. First, separate lexicons better accommodate bilingual processing of language pairs of different scripts (Kroll, van Hell, et al., 2010), e.g., Chinese and English, which were the L1 and L2 respectively of participants in the current study. Second, these three models have provided predictions on the effect of input language (i.e., L1 vs. L2) in bilingual lexicon development, which is the focus of the current study. The Bilingual Interactive Activation (the BIA; Dijkstra et al., 1998) and the BIA+ (Dijkstra & van Heuven, 2002) models did not make (explicit) predictions on the effect of input language.

The RHM (Figure 1) has three components, namely (1) the L1 lexicon, which stores the forms of L1 words, (2) the L2 lexicon, which contains the forms of L2 words, and (3) concepts, which are the meanings of words. Here, forms of words refer to the words' orthography, i.e., spelling and pronunciation. In the RHM, the forms of L1 and L2 words are stored separately and are connected via a lexical link (① in Figure 1). The forms of L1 words are connected to their meanings through direct and strong conceptual links (② in Figure 1). The connections between the forms of L2 words and meanings, on the other hand, are relatively weaker, especially when the words are newly learned and/or when learners are in the early stages of L2 learning. In this case, access to concepts for L2 words is usually mediated by the L1 through the lexical link. As L2 learners' proficiency increases, they may eventually be able to establish direct and strong conceptual links for L2 words (③ in Figure 1). Another key factor that influences learners' ability to directly access concepts for L2 words is the learning condition. van Hell and Kroll (2012) pointed out that a meaningful learning context, such as learning through pictures or real-life situations, contributed to the establishment of direct and strong conceptual links for L2

words while learning through L1 translations only strengthened the lexical links between L1 and L2 words instead of the direct links between L2 words and concepts.

Figure 1
Revised Hierarchical Model



Note. Adapted from “Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections between Bilingual Memory Representations”, by J. Kroll and E. Stuart, 1994, *Journal of Memory and Language*, 33, p.158

Jiang’s (2000) bilingual lexicon model elaborated on the developmental aspect of the RHM. Jiang proposed three stages for the development of L2 lexical knowledge. In the first stage, unlike an L1 word, which has L1-specific information for both form and meaning, an L2 lexical entry only has L2-specific information for form but not meaning. Instead, the L2 word entry contains a ‘pointer’ (p. 50) that links the L2 word to its corresponding L1 word. Access to meaning for the L2 word will be through its L1 counterpart via the pointer. Jiang hypothesized that the lack of L2-specific semantic information in the L2 word entry was partly due to the lack

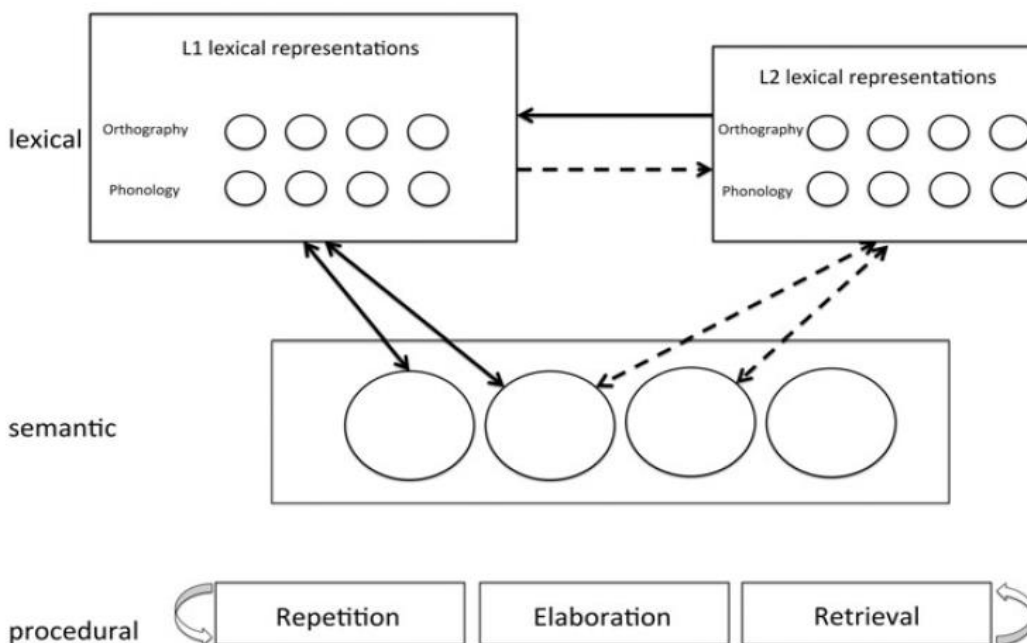
of attention to meaning during L2 word learning: learners can usually understand an L2 word through its L1 translation without the need to extract meaning from context. As one's exposure to the L2 increases, they may reach the second stage where L1 semantic information is integrated into an L2 word entry instead of being accessed via a pointer. Such integration means that activation of L1 translations for L2 words is now faster than when L1 translations are accessed via a pointer. What is common between the first and second stages is that access to meaning for an L2 word is mediated through the L1 and the link between the L2 word and its concept is weak. In the final stage of lexical development, an L2 word entry contains L2-specific information for both form and meaning and an L2 word is connected directly to its concept without L1 mediation. Jiang cautioned, however, that most L2 words may not reach the final stage partly because the words were initially taught through their L1 translations. Such learning method encourages reliance on the L1, and that subsequent exposure to the L2 words may only serve to reinforce this reliance, preventing the development of L2-specific semantics in a lexical entry.

The RHM-RER model (Rice & Tokowicz, 2020) focused on vocabulary training methods that would contribute to the establishment of direct conceptual links for L2 words. In the model, the L1 lexicon, L2 lexicon, and concepts were conceptualized as tiers (Figure 2). The first tier represents the forms of L1 and L2 words. The second tier contains the concepts of the words. Like in RHM, L1 words are connected to meanings directly whereas the connections between L2 words and meanings are mediated through the L1. Rice and Tokowicz added the third tier to illustrate training methods that strengthened the connections between the first tier (i.e., form) and the second tier (i.e., meaning), including repetition, elaboration, and retrieval. The authors stressed that the key to strong connections between the first and second tiers was to use training

methods that went across tiers, that is, to go beyond form-form connections between L1 and L2 words. Repetition, which usually involves repeating the L2 word and its L1 translation, may be effective but not sufficient to establish high-quality L2 lexical representations because repetition stays within the first tier, i.e., making L1-L2 form-form connections. Elaboration, on the other hand, encourages semantic processing by presenting a word in context, providing synonyms, or background information of the word. Putting the RHM-RER in the context of gloss language, L1 glosses that simply give L1 translations are similar to the repetition of L1-L2 word pairs while L2 glosses may contain synonyms and are similar to the method of elaboration.

Figure 2

Revised Hierarchical Model – Repetition, Elaboration, Retrieval



Note. From “A Review of Laboratory Studies of Adult Second Language Vocabulary Training”, by C.A.Rice and N. Tokowicz, 2020, *Studies in Second Language Acquisition*, 42, p.443 (<https://doi.org/10.1017/S0272263119000500>). Copyright 2019 by Cambridge University Press.

The prediction of the three abovementioned models that learning through the L1 might hinder the development of direct conceptual links for L2 words has been supported by a number of empirical studies in psycholinguistics in the context of intentional learning (e.g., Altarriba & Mathis, 1997; Comesaña et al., 2009; Elgort, 2011; Elgort & Piasecki, 2014; Finkbeiner & Nicol, 2003; Jeong et al., 2010). For example, Comesaña et al. (2009) taught Spanish-speaking children L2 Basque words through either L1 translations or pictures. The learners were then tested with a translation recognition task to see if they had established conceptual links for the newly learned words. Stimuli of the task were L1-L2 word pairs in three types of relations, namely translation, semantically related, and unrelated. The learners were asked to decide as fast as possible whether the L1-L2 word pair on the screen was a correct translation or not. The logic behind the task was that learners who had developed conceptual links for the L2 words would display the semantic interference effect, taking longer time and making more errors in rejecting semantically related pairs than unrelated one. Reaction time (RT) and accuracy data in the study revealed that learners who received picture explanations showed a greater semantic inference effect than those who learned through L1 translations, suggesting that learning through pictures was more conducive to the development of conceptual links for the L2 words than learning through L1 translations. The comparison of Elgort (2011) and Elgort and Piasecki (2014) offered more direct evidence regarding the effects of input language on word learning. Both studies followed similar learning and testing procedures: adult English L2 learners were first introduced to a set of pseudowords; learners were then given flashcards to take home to study the pseudowords before completing vocabulary posttests a week later. The flashcards provided the meanings of the pseudowords and an example sentence. The difference regarding the learning phase between the two studies was that Elgort (2011) used L2-only flashcards, which displayed the pseudowords along with their

explanations in the L2, while in Elgort and Piasecki (2014), the flashcards were bilingual, including pseudoword explanations in the L1. Both studies used a semantic priming task to assess semantic representations of the newly learned words. The idea was that only learners who had established conceptual links for the words would show semantic priming, i.e., reacting faster and more accurately to semantically related than unrelated word pairs. Results showed that in Elgort (2011), regardless of participants' proficiency level, those who learned through L2-only flashcards, displayed semantic priming, indicating that participants had developed direct links between L2 words and concepts. In Elgort and Piasecki (2014), only those with a larger L2 vocabulary size were able to do it. Based on the comparison of the two studies, Elgort and Piasecki (2014) concluded that L2-only flashcards led to lexical knowledge of higher-quality and were particularly beneficial for lower-proficiency learners.

Whether L2 words are connected to concepts directly or through L1 mediation carries real-world implications for L2 learning and use. van Hell and Kroll (2012) saw the ability to build direct links between L2 words and concepts a “hallmark” (p.154) of high proficiency. Rice and Tokowicz (2020) believed that the goal of L2 learning was being able to “think in L2” or “conceptually mediate the language” (p.455), which meant to bypass L1 mediation and access concepts directly in L2 word processing. Jiang (2000) expressed a similar opinion, arguing that direct conceptual links for L2 words contributed to fluent use of the words in communication whereas access to meaning through L1 was often effortful and lacked automaticity. Lexical fluency is considered an important aspect of word knowledge (Godfroid, 2019; Nation, 2001) and allows successful real-time language use (Schmitt, 2008). The importance of direct access to meanings for L2 words highlights the need to employ vocabulary teaching and learning approaches that promote the establishment of direct L2 form-meaning links. In the context of

gloss language research, the critical question is which type of glosses, i.e., L1 or L2 glosses, works more effectively for the development of such direct conceptual links and hence for greater lexical fluency. The psycholinguistic studies reviewed above had a more theoretical focus on the nature of lexical representations rather than measuring the actual processing fluency of word use. The current study directly examined how fluently learners were able to access L2 words in everyday tasks such as reading, which provides pedagogical implications on L2 vocabulary learning.

It is interesting to note that the RHM, Jiang (2000), and the RHM-RER have been interpreted differently with regard to their implications on the effects of input language. While psycholinguistics studies mostly deduced from the bilingual lexicon models that using L1 translations to learn L2 words would have negative effects, research on language of instruction in L2 classrooms and on gloss language tended to use the models to support L1 use, arguing that using L1 is beneficial, especially for lower-level learners, because these learners do not yet have strong connections between L2 words and concepts (e.g., Kang et al., 2020; Macaro & Lee, 2013; Zhao & Macaro, 2016). Rice and Tokowicz (2020) provided a more comprehensive interpretation of the RHM, Jiang (2000), and the RHM-RER models: there is more than one way to learn L2 words and learning through L1 translations may be easier for beginner learners (Kroll, van Hell, et al., 2010). Such interpretation highlights the roles of learners' individual differences, such as their L2 proficiency, in choosing input language. The current study, in comparing L1 and L2 glosses, took into account learners' L2 vocabulary size, among other factors, to provide a fuller picture of gloss language effect in vocabulary learning.

Gloss Language Research

As the review above on the use of L1 and L2 in L2 learning shows, findings in classroom studies on language of instruction have supported bilingual teaching, i.e., using both L1 and L2, while psycholinguistic research indicated that using L2 contributes to the development of direct conceptual links and hence lexical retrieval fluency for L2 words. The contradiction between these two lines of research warrants more research on the issue of L1 and L2 use. The current study attempted to investigate this issue from the perspective of gloss language, i.e., L1 versus L2 glosses, in vocabulary learning from reading.

There are several critical differences between gloss language research and each of those two lines of studies (classroom research on language of instruction and psycholinguistic studies on input language and bilingual lexicon). In terms of gloss language and language of instruction research, the language used in glosses is pre-planned, meaning that when it comes to L2 glosses, material writers can carefully select L2 words that are likely to be understood by learners of the targeted proficiency level. In comparison, teacher speech, which is often the focus of classroom research on language of instruction, is more spontaneous. Hence, teachers may not attend to whether all the L2 words they use can be understood by their students. Second, glosses are written input while teacher speech is spoken input. Written input is untimed and can be processed at learners' own pace; in contrast, spoken input is fleeting, requiring greater attentional resources on learners' part (K. M. Kim & Godfroid, 2019). Lastly, L2 glosses are short, compared to the continuous speech from teachers. It may be more challenging for learners, especially those in lower proficiency, to process such long speech in L2 whereas short L2 glosses are less likely to be an issue. The possible use of unfamiliar words, aggravated by the length and timed nature of spoken input, makes it hard for learners to comprehend and learn from L2

teacher speech, which may be the reason why L1 use is almost always advantageous over exclusive use of L2 in classroom studies. Regarding research on gloss language and on input language in psycholinguistics, the major difference is the learning condition: the former took place in incidental learning while the latter involved intentional learning. The difference in learning condition could lead to differences in learners' number of chances to see and engage with the target words.

In this section, I start by situating the use of glosses in the context of lexical focus on form and incidental vocabulary learning. I then present an overview of previous gloss language studies, before discussing potential variables that may moderate the comparative effectiveness of L1 and L2 glosses.

Glosses, Lexical Focus on Form, and Incidental Vocabulary Learning

Glosses are short definitions of words provided to support comprehension or word learning. Glossing belongs to an L2 instruction approach called lexical focus on form (e.g., Laufer, 2005, 2006; Laufer & Girsai, 2008). Focus on form takes place when learners' attention is briefly directed to linguistic forms, e.g., grammatical rules, in meaning-focused activities (Long, 1991, 1996). One way focus on form benefits learning is by allocating learners' limited attentional resources to both form and meaning. In meaning-focused activities, learners' primary attention is on meaning, e.g., reading comprehension and having conversations. Focus on form provides opportunities for learners to switch their attention temporarily to forms, which learners may otherwise not have the cognitive resources to attend to (Loewen, 2005, 2014; VanPatten, 1990). Most focus on form research has concerned the learning of morphosyntax (e.g., Ellis et al., 2006; Fu & Li, 2022; Sato & Loewen, 2018). When focus on form is applied to vocabulary

learning, i.e., lexical focus on form, the goal is to induce learners' attention to lexical items and to help learners establish accurate form-meaning connections for unfamiliar words.

Before further discussion on the effects of lexical focus on form and specifically, glossing, on vocabulary learning, it is necessary to make clear what vocabulary learning in meaning-focused activities is. According to Webb (2019), vocabulary learning in a meaning-focused activity, e.g., understanding the content rather than learning words, can be called incidental learning. Incidental learning is often thought of as 'picking up' words while doing something else, such as reading a book or watching TV, i.e., learning as a by-product of meaning-focused activities (Hulstijn, 2001; Loewen, 2014; Webb, 2019). Note that in such conceptualization of incidental learning, the core is the aim of the activity and not learners' behaviors, i.e., whether an activity is intended for vocabulary learning and not whether learners actually intend to learn new words in the activity. It is in fact hard to rule out intention to learn in incidental learning (Bruton et al., 2011; e.g., Hulstijn, 2001; Loewen, 2014). Several incidental vocabulary learning studies revealed that learners tried to memorize words encountered during reading (e.g., Y. Chen, 2021; Godfroid, Ahn, et al., 2018; Pellicer-Sánchez & Schmitt, 2010). The percentage of unknown words in the learning materials, the learning context (e.g., at home vs. in the classroom), and the use of typological enhancement are among many factors that may influence the presence and degrees of intention in an incidental learning condition (Webb, 2019). Ender (2016) used the term explicit processing to refer learners' attempts and strategies to learn new words, such as meaning inferencing and checking a dictionary, in incidental vocabulary learning. She argued that the use of these strategies, or the intention to learn, could exist in meaning-focused activities and did not alter the incidental nature of learning in these activities. In vocabulary research, besides making the activity a meaning-oriented one, incidental learning

can also be operationalized as a learning condition where learners are not forewarned of a posttest (Hulstijn, 2003). In the current study, the learning condition was incidental in that learners were asked to comprehend the reading instead of learning new words and they were not informed of the posttests after reading. Several terms have been used for incidental vocabulary learning in previous studies, such as contextual word learning (e.g., Elgort, Perfetti, et al., 2015; Elgort, Candry, et al., 2018; Elgort, Beliaeva, & Boers, 2020) and vocabulary learning from/during reading (e.g., Elgort & Warren, 2014). The current study use these terms interchangeably.

Going back to lexical focus on form, the reason why such approach is used is because while L2 learners are able to pick up words incidentally, the process is inefficient, taking considerable time while yielding limited gains. For example, in Godfroid, Ahn, et al. (2018), after encountering each target word in text around three times on average (range: 1–23 times), L2 learners learned scored around 30% in the form and meaning recognition posttests and 13% in the meaning recall test. Other studies showed similar percentages of gains in immediate posttests (e.g., 21% as measured by a meaning generation task in Elgort & Warren, 2014; 18% by a translation test in Waring & Takaki, 2003). In delayed posttests, gains dropped to 8% in a translation posttest a week after learning and to 4% three months later in Waring and Takaki (2003). Elgort and Warren (2014) suggested that at least 12 encounters with a word were required for noticeable learning. Eye tracking studies revealed that for novel words to be processed in a similar manner to familiar words, it took 10 exposures (Pellicer-Sánchez & Schmitt, 2010) or even over 40 (Elgort, Candry, et al., 2018).

The low learning gains in incidental conditions are mainly due to (1) lack of noticing of unfamiliar words, (2) inaccurate meaning inference, and (3) low retention of word knowledge

(Hulstijn et al., 1996; Laufer, 2005). Laufer (2005) elaborated on these reasons and argued that learners often overestimated their word knowledge and hence either failed to notice unfamiliar words or did not work on guessing word meanings; even when learners attempted to make guesses on the meanings of the new words, they did not always succeed because the context did not provide adequate clues; when learners correctly identified word meanings, they may still not be able to retain the meanings. Laufer (2005) maintained that incidental learning should not be the default approach for L2 learners and called for instructional intervention using lexical focus on form.

Types of lexical focus on form include glossing (e.g., Khezrlou et al., 2017; Warren et al., 2018), input enhancement, i.e., highlighting, bolding or underlining lexical items (e.g., Boers et al., 2017; Choi, 2017; Sonbul & Schmitt, 2013; Toomer & Elgort, 2019), and providing bimodal input (e.g., Y. Chen, 2021; Malone, 2018; Webb & Chang, 2015, 2022). Although input enhancement and bimodal input can increase the likelihood of learners noticing new vocabulary items, learners might still incorrectly infer word meanings, leading to erroneous form-meaning associations. Glosses enhance word salience while at the same time supply meanings, tackling two of the three major challenges learners face in incidental vocabulary learning, i.e., lack of noticing and erroneous meaning inference of unfamiliar words. According to the instance-based model for the learning of word meanings (Bolger et al., 2008), learners extract a word's core meaning, i.e., decontextualized as opposed to context-dependent word knowledge, through repeated encounters with the word in context; definitions given alongside context can accelerate this process by providing the core meaning directly.

Earlier research on glossing focused on the comparison between learning conditions with and without glosses (e.g., Hulstijn et al., 1996; Jacobs et al., 1994). In general, there is a large

positive effect of glossing on vocabulary learning, as suggested by a few meta-analyses (Abraham, 2008; Yanagisawa et al., 2020; Zhang & Ma, 2021; see Boers, 2022 for a review). Recent glossing research has witnessed increasing interest in the effects of gloss type, such as L1 versus L2 glosses (e.g., Choi, 2016; Kang et al., 2020; Ko, 2012) and multimodal versus text-only glosses (e.g., Boers et al., 2017; Jones, 2013; Ramezanali et al., 2021; Warren et al., 2018). The current study focused on gloss language, i.e., L1 versus L2 glosses. Gloss language is not only a relevant topic for language pedagogy, but also provides an interface to examine the implications of bilingual lexicon models, which make predictions regarding the consequences of learning through L1 and L2.

Previous Research on Gloss Language

Gloss language has been a contentious issue in vocabulary learning. From a pedagogical perspective, Laufer and Shmueli (1997) presented arguments both for and against the use of L1 and L2 glosses: on the one hand, students preferred L1 glosses, which allowed a sense of security about understanding the meaning of the words; on the other hand, learning through L1 translations may result in inaccurate uses of L2 words because there was not always a one-to-one correspondence between the L1 and the L2 for a given word (see also Jiang, 2000); further, L2 glosses provided additional exposure to the target language, which was believed to be beneficial for language learning. Findings from gloss language studies have been mixed. While many found L1 glosses to be more effective (e.g., Choi, 2016; Jacobs et al., 1994), others have demonstrated equal effectiveness (Kang et al., 2020; Ko, 2012; Yoshii, 2006) or superiority of L2 glosses (e.g., Miyasako, 2002; Shiki, 2008). Three recent meta-analyses on glossing (H. S. Kim et al., 2020; Yanagisawa et al., 2020; Zhang & Ma, 2021) also yielded contradicting findings. Yanagisawa et al. (2020), which focused on glossing in general, and H. S. Kim et al. (2020), which included

studies on gloss language only, both showed that L1 glosses were more effective than L2 ones in both immediate and delayed posttests, albeit H. S. Kim et al.'s comparisons had small effect sizes ($g = .44$ for immediate posttests; $g = .28$ for delayed posttests). Zhang and Ma (2021), in contrast, found that L2 glosses were more effective in the fixed-effect model, and the random-effect model showed no significant difference between L1 and L2 glosses. Note that many of the gloss language studies did not report whether words used in L2 glosses were familiar to learners. The different degrees of familiarity to words in L2 glosses in these studies may be one of the reasons for the inconclusive findings.

From a theoretical perspective, it follows from the Revised Hierarchical Model (RHM), Jiang (2000), and the Revised Hierarchical Model-Repetition Elaboration Retrieval model that L2 glosses might be more effective than L1 ones, at least in terms of establishing direct conceptual links for the L2 words. This prediction, though corroborated by some psycholinguistic research on intentional word learning, has not been consistently realized in research that compared L1 and L2 glosses in vocabulary learning from reading as shown by the abovementioned gloss language studies. The difference in learning condition, i.e., incidental in gloss language research and intentional in psycholinguistics studies, may have accounted for the discrepancy in results. In incidental vocabulary learning, learners are supported with context while in intentional learning, words are usually presented with less contextualization. For example, in Elgort (2011) and Elgort and Piasecki (2014), an example sentence was provided for each target word; in Jeong et al. (2010), target words were embedded in short videos showing real-world scenarios. In comparison, target words in gloss language research often appeared in a short story, which is much longer than a sentence or a short video. The rich context in gloss language research might have canceled out the negative effect of learning through the L1 and

allowed learners to connect L2 words with their concepts. Further, intentional learning means that learners can often review the target words as many times as they like while in most gloss language studies, learners only saw the target words and their definitions once. This means that learners in intentional learning can rehearse the link between a target word and its definition multiple times whereas those in incidental learning have only one opportunity to connect the target word and its gloss. It is somewhat expected that with only one exposure to the target words, glosses written in the L1, which is easier to process than the L2, provide a quick and easy way to establish form-meaning associations for words. Learners in the two learning conditions also differ in the amount of engagement with target words and word definitions. Learners are instructed to memorize words in an intentional condition while in incidental learning, the focus is on the comprehension of text where words are embedded, and learners often fail to notice unfamiliar words. In addition to the difference in learning condition, previous gloss language research assessed vocabulary gains using offline measures, which only offer insights into the product of cognitive processing. Psycholinguistic research, in contrast, used online measures that gauge the real-time lexical processing. It is possible that the processing differences between words learned through L1 and L2 glosses cannot not captured by offline measures alone. The prediction of the bilingual lexicon models that learning through L2 is more beneficial should be reevaluated in gloss language research with online outcome measures and with the number of target word encounters and learners' engagement with the words taken into account.

Potential Factors Moderating the Effects of Gloss Language

How many words learners are able to pick up in incidental learning is affected by a number of factors. In this study, I examined three, namely frequency of occurrence (FoO) of target words, i.e., the number of times a target word appears in text, learners' L2 vocabulary size,

and learner engagement. In this section, I focus on the first two factors, discussing how they affect incidental learning as shown by previous research and how they may moderate the effects of gloss language. Learner engagement, as a moderating variable and an outcome variable in this study, is reviewed in the next section.

Frequency of Occurrence. Vocabulary learning is an incremental process and the FoO of target words plays a critical role in the accumulation of lexical knowledge (Hulstijn, 2001). According to the instance-based models for word learning (Bolger et al., 2008; Reichle & Perfetti, 2003), each encounter with a word creates a memory trace that contains the word's form, meaning, and the context which the word is embedded in; the initial encounter only results in incomplete word knowledge; with each subsequent encounter, knowledge accumulated in previous encounters will be reactivated and eventually with sufficient experiences with the word, its meaning will be extracted. L2 incidental vocabulary learning studies have demonstrated such an incremental process empirically, showing that though learners were able to gain some lexical knowledge after one or two encounters with a word (e.g., C. Chen & Truscott, 2010; Malone, 2018), greater number of encounters, or higher FoOs of target words, led to better word learning (e.g., Godfroid, Ahn, et al., 2018; Vidal, 2011; Webb, 2007). Words of higher FoOs were more likely to be recognized and recalled in terms of both form and meaning as measured by offline posttests (e.g., Elgort & Warren, 2014; Webb, 2007; Webb & Chang, 2015); new words that are repeatedly encountered were also processed more fluently and in a similar manner to familiar words as assessed by online tests (e.g., Elgort, Beliaeva, & Boers, 2020; Godfroid, Ahn, et al., 2018; Pellicer-Sánchez, 2016; Pellicer-Sánchez & Schmitt, 2010).

FoO has been found to interact with lexical focus on form treatments in incidental vocabulary learning. In a meta-analysis on the effects of repetition in incidental word learning,

Uchihara et al. (2019) revealed an overall medium effect of FoO ($r = .34$), but the effect dropped when learning was assisted with multimodal input of reading-while-listening ($r = .28$) or viewing ($r = .22$), indicating that repetition might be more important in a ‘harsh’ unenhanced learning condition. Similarly, higher FoO may attenuate the effect of lexical focus on form. In Malone (2018), for example, there was a significant difference in vocabulary learning gains as measured by a form-recognition test between the reading-while-listening and reading-only conditions when target words appeared twice in text. However, such difference became nonsignificant when learners were exposed to target words four times.

The effects of FoO have rarely been researched in gloss research. Most studies on glossing included target words that appeared once in text. One exception, Teng (2020), supported the general trend that higher FoO led to better word learning. The study adopted a 2 (gloss vs. no gloss) x 3 (FoOs: 1, 3, & 7) between-subject design: learners saw 15 target words embedded in text once, three times, or seven times either with or without glosses. Posttests on recognition and recall revealed positive effects of FoO and glossing. There was also an interaction between FoO and glossing in that the effects of FoO were greater in the gloss than the no-gloss condition. Choi (2016), a gloss language study, found that L1 and L2 glosses were equally effective in the learning of target words that appeared twice as shown by both the immediate and delayed posttests; for target words with an FoO of four, L1 glosses worked better than L2 ones in the delayed but not in the immediate posttest. Given the mixed findings, further research is needed to clarify how FoO may influence the effects of glossing and gloss language.

L2 Vocabulary Size. L2 vocabulary size is a strong predictor of L2 proficiency (Qian & Lin, 2019). In this study, I used L2 vocabulary size as a proxy of L2 proficiency but will refrain from using these two terms interchangeably as proficiency is a multifaceted construct and is not

always equal to vocabulary size. Various other ways have been used to operationalize L2 proficiency in vocabulary studies, such as cloze test scores (e.g., Ko, 2012; Malone, 2018), automaticity in word retrieval (e.g., Elgort, Perfetti, et al., 2015; Elgort & Piasecki, 2014; Elgort & Warren, 2014), and academic status (e.g., Zhang & Ma, 2021). In what follows, I use the term ‘proficiency’ to refer broadly to learners’ language level regardless of how it is measured and save the term ‘vocabulary size’ for when learners are measured with vocabulary size tests.

In general, in an incidental condition, learners of higher proficiency gained greater word knowledge, as shown by their better performance in paper-and-pencil posttests (e.g., S. Lee & Pulido, 2017), ERP (e.g., Elgort, Perfetti, et al., 2015), and reaction time (RT; e.g., B. Chen et al., 2017; Elgort & Warren, 2014) data. Advanced learners also needed fewer encounters with target words to learn them (Uchihara et al., 2019). This is because higher-level learners may be able to better comprehend the text where target words are embedded and are thus more likely to successfully infer target word meanings based on context, even with few repetitions of the target words. The positive effect of higher proficiency can also be explained through a resonance mechanism, which hypothesized that known words are nonselectively activated when relevant words are read (Myers & O’Brien, 1998). It follows that in incidental word learning, advanced learners have more known words to be activated when reading a target word, leading to faster establishment of stronger connections between target words and existing words. When glosses were provided, however, proficiency did not seem to moderate learning gains, as shown by Yanagisawa et al.’s (2020) meta-analysis, which was probably due to the lack of need to infer meanings.

When it comes to proficiency and gloss language, it is logical to conjecture that L2 glosses are less effective for lower-proficiency than for higher-proficiency learners. Finding L2

glosses challenging to understand, learners with limited proficiency are more likely to misunderstand the glosses or simply to ignore them (Boers, 2022). Results regarding the role of L2 proficiency in gloss language, however, have been mixed. Yanagisawa et al. (2020) and Zhang and Ma (2021)'s meta-analyses on glossing did not find a moderating effect of L2 proficiency on gloss language. In contrast, H. S. Kim et al. (2020)'s meta-analysis on gloss language found that L1 glosses were more effective for lower-level learners. Inconsistency in findings may be due to different operationalizations of L2 proficiency and also outcome measures of vocabulary gains. More research is needed to understand the role of L2 proficiency in the comparison of L1 and L2 glosses, particularly research that takes into consideration other variables, e.g., target word FoO and learners' engagement with glosses, and uses various other outcome measures. Recall that in Elgort and Piasecki (2014), when being able to view target words multiple times and being measured with a RT-based semantic priming task, learners of lower proficiency benefitted more from L2 than from L1 word definitions, opposite to findings in H. S. Kim et al. (2020).

Learner Engagement in Vocabulary Learning

Defining Engagement

Although engagement is a common term used in everyday life, it can be an elusive concept and encompasses many different phenomena. Here, I differentiate between two types of engagement, engagement as attention and engagement as action. Engagement as attention simply means paying attention to something, e.g., noticing that a word is unfamiliar. Engagement in action goes beyond mere noticing and refers to actions taken on something, e.g., looking up a word in a dictionary. Engagement as action is similar to the concept of involvement in the Involvement Load Hypothesis for vocabulary learning (Hulstijn & Laufer, 2001; Laufer &

Hulstijn, 2001). According to the hypothesis, involvement has three components, namely need, search, and evaluation. Need concerns how motivated a learner is to do a certain thing in order to complete a task. The need to check a word's pronunciation is strong when a learner wants to use the word in a speaking task. Search is the action taken to figure out the meaning of an unfamiliar word, such as asking instructors or peers about the word. Evaluation refers to learners' assessment of how the word fits in its context. Engagement as action depends on engagement as attention, i.e., attention is the prerequisite of action. It is hard to imagine a learner will look up a word if they haven't noticed the word first. In the Involvement Load Hypothesis, search and evaluation are learners' actions on unfamiliar words and they are "contingent upon allocating attention to form-meaning-relationships" (Hulstijn & Laufer, 2001, p. 543). Engagement as attention, in contrast, can take place without engagement as action. A learner may notice an unknown word but decide not to do anything with it.

Measuring Engagement

Engagement in vocabulary learning studies has been gauged with eye tracking, think alouds, retrospective surveys, and tracking learner behaviors in computer- and mobile- assisted language learning. Except for retrospective surveys, the other three methods are able to reveal real-time engagement as learners are processing the learning materials. In what follows, I review how these methods measure engagement as action and as attention.

Eye tracking is mostly used to measure engagement as attention. In many studies, total reading time on a word was used as an index of attention paid to the word (e.g., Godfroid, Boers, & Housen, 2013; Godfroid, Ahn, et al., 2018; Mohamed, 2018; Pellicer-Sánchez, 2016). Eye tracking can also reveal engagement as action when the actions are performed on screen. Warren et al. (2018) examined eye movements to target words and three types of marginal glosses (text,

picture, and multimodal). In this study, eye movements towards target words can be seen as engagement as attention; eye movements towards marginal glosses indicated engagement as action: learners took actions to look away from the main text to page margins to consult glosses. However, eye tracking in this study may not be able to reveal learners' other actions during reading, such as inferring word meanings.

Think-alouds probe concurrent processing by asking participants to articulate their thoughts while doing something. Ender (2016) used this method to examine the cognitive processes during incidental vocabulary learning. Based on the think-alouds data, she categorized participants' processing strategies into ignoring a word, checking a dictionary, inferring meaning from context, and inferring meaning plus checking a dictionary. Ignoring indicated engagement as attention because this category included instances where participants would notice that a word was unfamiliar but decide not to consult a dictionary. The other three categories can be seen as engagement as action. Ender treated uncommented words as unattended ones, i.e., without engagement as attention. However, it was hard to ascertain whether participants underwent unarticulated processing of those words.

In retrospective surveys on engagement, learners self-report whether they pay attention to unfamiliar words and what they do with the words. Elgort and Warren (2014), for example, used a five-point Likert scale to examine the extent to which learners (1) ignored unfamiliar words, (2) tried to infer the words' meanings, and (3) noted down the words and checked the dictionary later. Like in think-alouds, ignoring here would mean engagement as attention without action and the other two options would indicate engagement as action. A limitation of retrospective surveys is that they can only give a general picture of learners' thoughts during learning but are not able to tell us how learners process each target word.

When learning takes place on a computer or a mobile device, learners' behaviors such as mouse clicks, mouse movements, and keystrokes can be recorded (see Fischer, 2007 for a review on tracking in computer-assisted language learning). Vocabulary learning studies have mostly used tracking data to examine whether and how learners used glosses and online dictionaries (e.g., Chun & Payne, 2004; Laufer & Hill, 2000; H. Lee et al., 2017; Peters, 2007; Varol & Erçetin, 2021). Tracking data can reveal the number of times, the frequency, and the duration of gloss access and dictionary lookups. Tracking data can thus be used to indicate engagement as action, e.g., what learners do with glosses, but may not be able to say a lot about engagement as attention — when learners do not click a gloss, they may still have paid attention to the word. Compared with think-alouds, tracking has the advantage of being unobtrusive. Tracking on a computer or a phone is more ecologically valid than eye tracking because the former allows learners to perform a task anywhere anytime as long as they have a digital device where the tracking program is implemented while the latter mostly requires learners to sit in the lab, which is not a typical environment learning takes place in reality.

Engagement as a Moderating Factor: What are the Effects of Engagement?

It is a “commonsense notion” (Schmitt, 2008, p. 338) that more engagement predicts better learning. Craik and Lockhart's (1972) Levels of Processing framework is often cited to support this notion. According to the framework, analysis of stimuli goes from the shallow processing of physical forms to the deeper levels of meaning extraction and elaboration. Deeper processing leads to stronger and longer-lasting memory traces. The Involvement Load Hypothesis mentioned above also argued that higher involvement resulted in greater word knowledge. The positive relationship between engagement and vocabulary learning is evidenced in empirical research. Eye-tracking studies have shown that longer reading times, an indication

of greater engagement as attention, resulted in higher vocabulary gains in incidental learning (e.g., Godfroid, Ahn, et al., 2018; Pellicer-Sánchez, 2016). Greater engagement as action also leads to better learning. In Ender (2016), for example, only 12% of the ignored target words were recalled in posttests, compared to 27% for words that were looked up in a dictionary. Engagement as action not only leads to greater vocabulary gains but also accelerates the learning process. In Elgort and Warren (2014), learners who tried to infer word meanings needed fewer encounters with a word to learn it. Similarly, Uchihara et al.'s (2019) meta-analysis on repetition in vocabulary learning revealed that the effects of target word FoO diminished when learners used dictionaries, asked questions about words, or took notes.

In research on glossing, consulting a gloss can be seen as engagement as action. The amount of engagement in gloss research was usually operationalized as the number of times learners accessed glosses and the duration learners spent reading the glosses. Findings showed that the relationship between learners' engagement with glosses and learning gains was not straightforward. Laufer and Hill (2000) found no relationship between the number of times learners accessed word definitions and word knowledge measured by a meaning recall test. Warren et al. (2018) used eye tracking to measure engagement as action, i.e., learners' reading time on glosses. The authors, like Laufer and Hill (2000), found no relationship between time on gloss and learning. H. Lee et al. (2017)'s results revealed a negative correlation between the amount of time learners spent on glosses and learning gains and a non-significant correlation between the frequency of gloss access and learning. Finally, in Peters (2007), greater number of gloss access led to more learning. This positive relationship was moderated by whether a target word was relevant to the comprehension questions: the correlation between gloss access and learning was higher for relevant than for nonrelevant words. The discrepancy in results in these

studies is likely due to various reasons. As Peters (2007) suggested, relevance of a target word could be one of the reasons. Warren et al. (2018) indicated that gloss type could be another. While the study did not find a relationship between time on glosses and learning, it showed that time spent on target words and picture glosses together yielded higher gains than text and multimodal glosses, which the authors attributed to the mnemonic advantage of picture glosses. No gloss language research, to my knowledge, has examined how engagement with L1 and L2 glosses may be differentially associated with learning. Such examination will advance our understanding of both engagement and gloss language in vocabulary learning.

Engagement as an Outcome Variable: What Affects Engagement?

Engagement may be affected by the learning condition and learners' characteristics. The Involvement Load Hypothesis (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001) predicted the engagement level of a learning condition based on whether the condition instigated learners' need, search, and evaluation. For example, engagement is higher when learners are required to select the correct word meaning from several options than when they are given the meaning. In glossing research, many studies have found that types of glosses influenced learner engagement (Rassaei, 2020; Türk & Erçetin, 2014; Varol & Erçetin, 2021; Warren et al., 2018; cf. H. Lee et al., 2017). Warren et al. (2018) showed that text glosses were mostly likely to be ignored, followed by picture and multimodal glosses, though there was no significant difference in attention paid to the three types of glosses. In terms of learners' characteristics, Chun and Payne (2004) found that lower working memory was associated with more gloss lookups, i.e., greater engagement as action. Varol and Erçetin (2021) explored how gloss type and working memory interactively affected engagement with glosses. Participants in the study read with one of the four types of glosses that differed in either content (lexical vs. topical) or position (pop-up vs.

separate windows). Findings revealed no interaction between working memory and gloss type, with the frequency of gloss access being affected by gloss position only. Gloss language studies have rarely looked into how learners engage with L1 and L2 glosses. Results in Chun and Payne (2004) suggested that when learners had the freedom to pick, they accessed L1 glosses more frequently than L2 glosses. Laufer and Hill (2000) alluded to the possibility that learner characteristics may influence preference for L1 or L2 glosses. In the study, while Israeli high school students accessed L1 glosses more, college students in Hong Kong preferred L2 glosses. Given that the learner characteristic of L2 proficiency has been found to influence the amount of and learner attitudes towards L1 use in L2 classrooms (e.g., DiCamilla & Antón, 2012; J. H. Lee & Lo, 2017), it is possible that proficiency level also affects how much learners engage with L1 and L2 glosses.

The Present Study

The goal of the study was to unpack factors that may interactively and independently moderate gloss language effect on learners' engagement with glosses and on vocabulary learning from reading. Such endeavor contributes to our knowledge of when to use L1 and L2 glosses to optimize learning and acknowledges the critical role of L1 in L2 learning. Specifically, I examined how learners' L2 vocabulary size, engagement with glosses, and target words' FoO affected the comparative effectiveness of L1 and L2 glosses in vocabulary learning and retention. I also explored how learners' vocabulary size influenced their engagement with L1 and L2 glosses. Because a target word was glossed only at its first occurrence and learners did not know beforehand how many times a target word would appear in text, FoO was not likely to affect learners' engagement with the first encounter of the target word and its gloss. Learning was measured in terms of both receptive and productive meaning knowledge, and fluency of word

retrieval. The assessment of target words' retrieval fluency went beyond most previous gloss language research and was able to shed some indirect light on the prediction of the RHM that L2 input facilitates the establishment of direct conceptual links for L2 words and thus more fluent word retrieval. Both the short-term and long-term development of word knowledge was measured, by immediate and delayed posttests respectively. The operationalization of variables involved in the study is summarized in Table 1.

Table 1
Operationalization of Variables

Variables	Operationalization	Measures
Gloss engagement	Time on gloss	Computer logs
Vocabulary size	Receptive vocabulary size	Updated Vocabulary Levels Test
Receptive meaning knowledge	Accuracy of meaning matching	Meaning matching test
Productive meaning knowledge	Accuracy of meaning recall	Meaning recall test
Fluency of lexical retrieval	Reaction time	Self-paced reading test
Short-term development	Accuracy of immediate posttests	Immediate posttests
Long-term development	Accuracy of delayed posttests	Delayed posttests

The following research questions guided the study:

RQ1. How does gloss language affect learners' engagement with glosses, as moderated by learners' vocabulary size?

RQ2. How does gloss language affect learning in short-term and long-term receptive and productive meaning knowledge, and lexical retrieval fluency, as moderated by target word FoO, learners' vocabulary size, and learner engagement?

Previous research on the effects of FoO, engagement, and L2 proficiency justified several possible interactions between the variables being examined. For RQ1, the interaction between gloss language and vocabulary size is plausible. For RQ2, the possible interactions include (1) FoO and gloss language, (2) vocabulary size and gloss language, (3) gloss engagement and gloss language, (4) FoO, vocabulary size, and gloss language, (5) FoO, gloss engagement, and gloss language, and (6) vocabulary size, gloss engagement, and gloss language.

CHAPTER: 2 METHOD

This chapter presents methodological details regarding participants, materials, procedure, test scoring approach, and data analysis. Instructions participant received for each task, i.e., reading, exit questionnaire, language background questionnaire, and the three vocabulary posttests, are included in Appendix A. All instructions were written in participants' first language (L1).

Participants

One hundred and eighteen second language (L2) learners of English completed all the tasks in the study. The participants were recruited through word-of-mouth and flyers posted on Wechat, a social media platform widely used in China. Participants were randomly assigned to either the L1 gloss or the L2 gloss group. Eight participants were excluded from data analysis due to low reading comprehension score (see Reading Comprehension in the Data Analysis section). The final sample size included in data analysis was 60 in the L1 gloss group and 50 in the L2 gloss group.

All included participants spoke Chinese as their L1 and were studying at university in China at the time of participation. Participants came from different institutions. Their mean age was 20.93 years ($SD = 2.20$, range: 17–30). On average, the participants had had English classroom instruction in China for 11.59 years ($SD = 2.36$, range: 6–18) and were at different levels of degree programs, e.g., undergraduate ($n = 91$), master's ($n = 18$), and PhD ($n = 1$).

All participants took the updated Vocabulary Levels Test (Webb et al., 2017; see more details about the test in the section on L2 vocabulary size) and received a score of 20 or above for each of the 1000, 2000, and 3000 levels. The threshold of 20 indicated that the participants had a mastery of words in the three levels (Nation, 1983; but see Webb et al., 2017) and that thus

participants were not likely to have difficulty comprehending the reading material and the glosses (see more details in the Materials section below). Participants' average score in the updated Vocabulary Levels Test was 134.90 ($SD = 11.03$; range: 102 – 150) out of a maximum score of 150, suggesting that the participants were of intermediate proficiency level. The Cronbach's α reliability estimate of the updated Vocabulary Levels Test was .91.

Materials

Reading Material

I adapted the introduction and the first 13 chapters of the Pearson graded reader *The Client* to be the reading material. The graded reader is a thriller, and the suspense in the story may help engage learners and keep their interest in the content. The purpose of the adaptation was to increase some of the target words' frequency of occurrence (FoO). The reading has 9,808 words. An analysis by Vocabprofile on Lextutor (Cobb, n.d.) revealed that 98% of the words in the reading are from the most frequent 2,000 word families in the BNC/COCA word frequency list (Nation, 2012). See Appendix B for a complete vocabulary profile of the reading material.

Target Words

Twenty-four words with an FoO ranging from one to 27 ($M = 6.58$, $SD = 6.13$) in the reading material were replaced by pseudowords to avoid out-of-experiment exposure and prior knowledge. The 24 words were chosen because of their wide range of FoOs. The target words were underlined, bolded, and colored in blue to indicate the availability of glosses. The pseudowords' parts-of-speech remained unchanged from the original words, which included 17 nouns, four verbs and three adjectives. The ratio of target words was 1.61% in text.

The pseudowords were selected from the English Lexicon Project (Balota et al., 2007), a corpus that contains behavioral data, i.e., reaction time (RT) and accuracy, of 40,481 words and

40,481 pseudowords from 816 L1 speakers of English in a lexical decision task and 444 in a naming task. The pseudowords in the project were generated by changing one or two letters in real words (Balota et al., 2007). The pseudowords were between four and seven letters long ($M = 5.5$, $SD = .72$). This range of length was chosen because pseudowords that are one or two letters different from real words are more word-like when they are longer (Keuleers & Brysbaert, 2010). The chosen pseudowords also had similar RTs as obtained from the English Lexicon Project ($M = 842.23$; $SD = 25.95$). Because the degree of a pseudoword's resemblance to a real word affects its RT (Keuleers & Brysbaert, 2010), comparability in RTs indicated that the chosen pseudowords were similar in their word-likeness, which reduced the likelihood that some pseudowords were more salient and were easier to learn than others (Bartolotti & Marian, 2017). The pseudowords were also comparable in the number of orthographic neighbors and mean bigram frequency, which are believed to affect word processing. Appendix C presents the selected pseudowords, their characteristics obtained from the English Lexicon Project, and the original words they replaced.

Glosses

Two types of glosses were constructed, namely L2 and L1 glosses. L2 glosses were modifying definitions of the corresponding real words. L1 glosses contained the L1 translations of the L2 glosses. To make the L2 glosses comprehensible, all words used in the L2 glosses were within the most frequent 3,000 word families in Nation's (2012) BNC/COCA word frequency list. The two types of glosses were matched in length: L2 glosses were 4.17 words long ($SD = 1.83$) and L1 glosses 4.54 ($SD = 2.17$) characters long on average.

The gloss of a target word appeared in a pop-up window (see more details in the Reading Interface section) when participants clicked the target word. The pop-up window closed when

participants clicked the target word again. Each target word was glossed once on its first occurrence. Examples for L1 and L2 glosses are provided below. For the full list of glosses, see Appendix D.

Example: glosses for the target word ‘haron’

L1 gloss: 重要政治人物

L2 gloss: an important politician

Reading Comprehension Questions

Ten comprehension questions were inserted in the reading material with an interval of about six pages. The comprehension questions were taken from the Pearson teachers’ resources for the graded reader and were all closed-ended questions, i.e., multiple choice and true/false items.

The Reading Interface

Figure 3 is an example of the reading interface as shown on a laptop. The reading material was presented across 25 pages. Each page contained around 400 words, except for the first page, where the introduction part of the reading was presented (132 words), and the last page (249 words). The text of the reading was in Times New Roman font in 18 pixels in the color of black. The text of the glosses was in the same font in 15 pixels in white against a black background. Gloss windows always popped up below target words. Page numbers and the total pages were displayed on the bottom left corner. Page numbers started from the first page of the reading or the first page after comprehension questions (whichever applied). The total pages represented the number of pages participants had to read in total before encountering a comprehension question. For example, in Figure 3, page 3/5 means that participants are on the third page of the reading or the third page after comprehension questions, and there are five

pages in total before the next set of comprehension questions come up. Below the page numbers, the ‘next page’ button allowed participants to proceed to the next screen. Participants could not go back to previous pages once they click the ‘next page’ button.

Figure 3

Example of the Reading Interface

'It's because of the hoag,' said the man. 'It's at my house. I'm a lawyer. My client killed a [haron](#). The FBI. It is looking everywhere for his hoag and my client hid it in my [goncho](#).'

Mark sme a room to store things now, thanks to Ricky.

'Who's your client, Romey?' he asked.

'Barry the Blade,' said Romey. 'He's a member of the valoon. Now he wants to kill me because I know about the hoag. But he can't because the gas will kill us first.' He laughed, and drank more vandier.

Soon he was drunk and asleep. Mark gently opened the car door and ran to where Ricky was hiding.

'I pulled the tube out,' said Ricky in a high voice.

page 3 / 5

Next Page

Exit Questionnaire

The exit questionnaire (see Appendix E), written in participants' L1, had 10 questions, and was administered primarily to probe participants' gloss access. The first question asked participants how often they checked glosses. In another question, participants indicated to what extent they skipped a gloss for the following four reasons: (1) they have guessed the word meanings; (2) they did not need the gloss to understand the reading; (3) knowing a word's meaning was not important; and (4) the gloss was not helpful. In these two questions, participants gave their responses on a scale from 0 to 100 for these questions. 0 meant that participants did not check glosses at all, and 100 meant that participants checked all the glosses in the first question. In the second question, 0 indicated that a given reason was never why participants skipped glosses, and 100 indicated that a particular reason was always why

participants skipped glosses. Participants were also prompted to give examples of glosses that had helped them with reading comprehension or word learning.

The exit questionnaire also included questions about whether participants understood the glosses, whether the glosses had helped with reading comprehension and word learning, and how much participants enjoyed the reading. Responses to these questions were made on a 100-point scale. Participants also indicated whether they had deliberately memorized the target words and whether they had guessed that there would be vocabulary posttests. These two questions required yes or no answers.

Language Background Questionnaire

The language background questionnaire (see Appendix F) was administered to collect details about their language learning history and use. There were three sections in the questionnaire. The first section asked participants to provide basic information, including their name, participant ID, age, gender, and education level. The second part focused on participants' language proficiency. It asked participants to self-assess their overall English proficiency as well as their proficiency in reading, listening, writing, and speaking on a 10-point Likert scale. In this section, participants also provided their scores of standardized tests they had taken (e.g., TOEFL and IELTS). The last part looked into learners' English learning history and current English use: age of onset of English learning, ways of learning, years of formal language education in the classroom, current amount of English classroom instruction (hours per week), and experiences of living or studying in an English-speaking country.

L2 Vocabulary Size

Participants' vocabulary size was measured by the updated Vocabulary Levels Test (Webb et al., 2017). The test assesses learners' receptive word knowledge at five word frequency

levels from 1,000 to 5,000. Each level has 10 clusters, with six words (three target items and three distractors) and three word meanings in each cluster (see Figure 4 for an example). Each cluster is worth three points and each level had a maximum score of 30 points. The maximum score of this test is 150. In each cluster, test takers are asked to match the words on the right with the word meanings of the left. Two equivalent versions were originally created, and Form A was used in the current study (see Webb et al., 2017, Appendix 1).

Figure 4
Example of the updated Vocabulary Levels Test

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

Vocabulary Posttests

I tested participants' vocabulary knowledge of the target words with a meaning recall test, a meaning matching test, and a self-paced reading test. The tests were administered immediately after reading (i.e., immediate posttests) and two weeks after the second reading session (i.e., delayed posttests). Each test examined a different aspect of word knowledge: the meaning recall and matching tests assessed participants' ability to productively retrieve and receptively recognize word meanings, respectively, and the self-paced reading test tapped into how participants retrieve and integrate meaning in real-time reading.

In the meaning recall test, participants saw a target word and were asked to type its L1 translation equivalent, L2 synonym, or a short definition in either the L1 or the L2. Example test items are presented in Figure 5. In the meaning matching test (see Figure 6), participants saw the 24 target words on the second half of the screen. The target words' L1 and L2 glosses, along

with two additional word definitions serving as distractors, were presented on the first half. Each gloss was numbered, and participants were asked to match the target words with their meanings by selecting from a drop-down menu the number of the corresponding gloss. For both the meaning recall and the meaning matching tests, response to each item was mandatory. Participants were instructed to type a question mark“?” when they did not know the answer in the meaning recall test.

Figure 5
Example of the meaning recall test

latpin

valoon

corax

haron

caudam

Figure 6*Example of the meaning matching test*

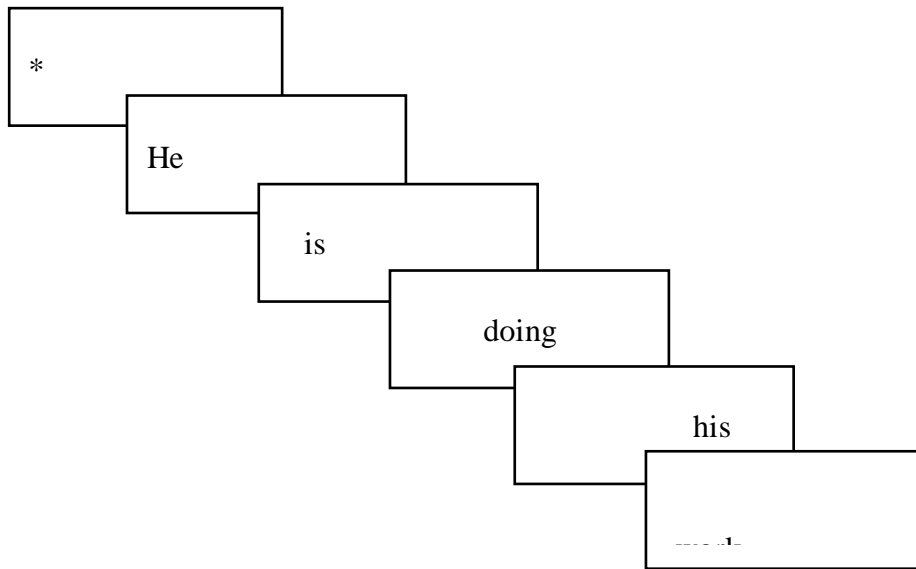
1 police not in uniform 便衣警察	14 where car smoke gets out 汽车排气管
2 a criminal organization 犯罪集团	15 a room to store things 杂物存储间
3 an important politician 重要政治人物	16 strong alcohol 烈酒
4 a car used as a house for people to live 供人居住的改装车	17 a drink without alcohol 不含酒精的饮料



For the self-paced reading test, I adopted the moving window paradigm (see Figure 7 for an illustration), which is most widely used and yields results the best correlate with gaze durations in eye tracking (Jiang, 2013; Just et al., 1982). Each trial in the self-paced reading test began with an asterisk. The asterisk represented the position where a sentence starts. Participants

proceeded through a trial by pressing the space bar. With each press, a word appeared. In a moving window paradigm, a subsequent word appears to the right of the preceding word, and the preceding word disappears upon the presentation of the subsequent word.

Figure 7
The moving window paradigm



24 context-neutral sentences were constructed for the self-paced reading test. Three versions were created for a sentence, each containing a target word, a real word, or a nonword at the same position, i.e., the critical position. The critical position and the two words that follow were never in the sentence-final position to avoid the sentence wrap-up effect, i.e., longer RT to the last word in a sentence. The nonwords used in the self-paced task, like the target words, were obtained from the English Lexicon Project. Three counterbalanced lists were constructed such that each version of a sentence appeared once in each list and all three conditions appeared across the three lists (see Table 2). Trials were randomized. All of the sentences were followed by a comprehension question to encourage participants to focus on reading for meaning. Stimuli

characteristics obtained from the English Lexicon Project were summarized in Table 3. Kruskal-Wallis tests suggested that there was no significant difference between pseudowords, nonwords, and real words in length ($\chi^2(2) = 2.79, p = .25, \epsilon^2 = .04$), in the number of orthographic neighbors ($\chi^2(2) = 2.59, p = .27, \epsilon^2 = .04$), and in mean bigram frequency ($\chi^2(2) = 3.24, p = .20, \epsilon^2 = .05$). There was no significant difference between pseudowords and nonwords in their mean RTs in the English Lexicon Project as suggested by a Mann Whitney U test ($W = 374.00, p = .08, r = .26$). For a list of full stimuli in the self-paced reading test see Appendix G.

Table 2
Examples of Self-Paced Reading Stimuli

Sentence	List 1	List 2	List 3
Jason saw a <u>(critical position)</u> in front of the shop.	latpin (pseudoword)	police (real word)	royate (nonword)
They talked about the <u>(critical position)</u> very often during dinner.	remude (nonword)	valoon (pseudoword)	student (real word)
He took a photo for the <u>(critical position)</u> and his wife.	teacher (real word)	persn (nonword)	haron (pseudoword)

Table 3
Stimuli Characteristics

Stimuli type	Length (letters)	Orthographic neighbor	Mean bigram frequency	Mean RT (ms)
Pseudowords	5.50 (.72)	1.83 (1.13)	1794.13 (826.40)	842.23 (25.95)
Nonwords	5.92 (.72)	1.33 (.76)	2028.95 (836.54)	859.89 (29.54)
Real words	5.62 (1.31)	3.04 (3.44)	1577.85 (679.04)	NA

Procedure

Participants were first screened by the updated Vocabulary Levels Test and the language background questionnaire. Those who passed the 1000, 2000, and 3000 levels, i.e., scoring above 66% in each level (Nation, 1983), and who were college or graduate students in China continued with subsequent tasks. One hundred and eighteen out of the 216 who participated in the screening procedure were eligible to continue their participation. Eligible participants were randomly assigned to read with L1 glosses (i.e., the L1 gloss group) or with L2 glosses (i.e., the L2 gloss group). Given the length of the reading, I administered two reading sessions on two consecutive days. Participants read 18 out of the 25 pages in the first session and the remaining seven pages in the second session. Participants could choose to read on a laptop, a tablet, or a phone. Reading instructions were presented before each session. The reading sessions were untimed and self-paced.

After the completion of the second reading session, participants completed the exit questionnaire, which was followed by surprise (i.e., unannounced beforehand) immediate

vocabulary posttests in the order of a self-paced reading test, a meaning recall test, and a meaning matching test. Delayed posttests were administered two weeks after the second reading session. The total duration of the study was around 3.5 hours. Participants were asked not to discuss the study with other people to prevent friends who also participated from knowing about the posttests. In a debriefing at the end of the study, I offered the .txt version of the graded reader and discussed the design of the study with participants who were interested. Participants were given 100 RMB (around 15 US dollars) after they completed the study as compensation for their time.

The study was implemented online. Data were collected using Qualtrics and Gorilla Online Experiment Builder. Table 4 presents the online platforms used and estimated duration for each task in the study.

Table 4*Study Timeline, Task Duration, and Data Collection Platforms*

	Task	Duration	Platform
Day 1	Updated Vocabulary Levels Test	30 minutes	Qualtrics
Day 2	Reading session I	90 minutes	Gorilla
Day 3	Reading session II	20 minutes	Gorilla
	Exit questionnaire	5 minutes	Gorilla
	Self-paced reading test	10 minutes	Gorilla
	Meaning recall test	10 minutes	Gorilla
	Meaning matching test	10 minutes	Gorilla
Day 17	Delayed posttests (self-paced reading, meaning recall, meaning matching)	30 minutes	Gorilla
Total		205 minutes	

Data Analysis

In this section, I first provide details about the coding of responses and data trimming procedures for each instrument used in the study. I then describe the data analysis approach I used to model the data.

Reading Comprehension

One point was awarded for each correct response to the reading comprehension questions, with 10 points as the maximum score. The average score for the reading comprehension is 7.65 ($SD = 1.45$; range: 1–10), demonstrating adequate comprehension of the reading material by the majority of participants (see Elgort & Warren, 2014; Godfroid, Ahn, et

al., 2018). Eight participants were excluded from data analysis because they had a reading comprehension score below 6, which indicated that these participants might not have fully understood the reading material or were not paying sufficient attention to meaning during reading.

Time on Gloss

Time on gloss was recorded as the time difference between when a participant clicked open a gloss and when a participant clicked the gloss again to close the gloss pop-up window. Time on gloss referred to the total time spent on one particular gloss. If participants clicked a gloss multiple times, the time on that gloss would be the summed duration of each click. The original range of gloss time was between 354 milliseconds (ms) and 2234172 ms (around 37 minutes). The extreme values indicated that the recorded gloss time may not have accurately reflected participants' actual engagement time with the gloss, i.e., time that participants actually spent on reading the gloss. Data cleaning was needed in this case to exclude or adjust gloss time values that did not reflect participants cognitive processes involved in reading a gloss.

The cutoffs for data cleaning were determined based on research on reading rates and research using RT tasks. In a review and meta-analysis on reading rates, Brysbaert (2019) estimated that the reading rates of L1 speakers of English were around 238 words per minute for nonfictions (around 250 ms a word) and 260 words per minute for fictions (around 230 ms a word). L2 speakers read slower, with a reading speed ranging from 139 to 174 words per minute (around 344 to 430 ms a word). In self-paced reading and lexical decision tasks, the lower cutoff of RT cleaning is usually set between 100 and 250 ms and the upper cutoff between 2500 and 3000 ms per word (Jiang, 2013; Marsden et al., 2018). With the goal to preserve as many data points as possible, i.e., minimal data cleaning, 250 and 3000 ms were chosen as the lower and

upper cutoffs to identify outliers in the gloss time data: 250 ms per word was not as slow as the reading rates of L2 speakers, yet not the fastest reading rates of L1 speakers; and 3000 ms was at the higher end of the typical upper cutoffs in RT research. Further outliers would be taken care of in model criticism (see the Mixed-effects Modelling section). Specifically, the upper cutoff for each gloss was calculated as the number of words in the gloss multiplied by 3000 ms and the lower cutoff as the number of words in a gloss multiplied by 250 ms. Values of time on a gloss above the upper cutoff for that gloss were winsorized and replaced by the upper cutoff of that gloss. Winsorization instead of trimming was used because extremely long gloss time still reflected some cognitive processing of a gloss and thus should not be completely removed. Such data cleaning procedure affected 20.64% of the data. Values of time on a gloss below the lower cutoff were replaced by zero. This was because participants were unlikely to have processed the gloss within such a short time. This data cleaning procedure affected 11.59% of the data.

Meaning Recall Tests

Correct or partially correct answers were coded as 1 and others (e.g., incorrect answers and blanks) as 0. Correct answers were defined as ones that included all the semantic features of a target word. Partially correct answers either (1) contained some but not all of the semantic features or (2) included all the semantic features plus some features that were not in the word meaning. Table 5 gives examples of correct and partially correct answers as determined by the above method. The total data points for each participant were 24. The reliability estimates (Cronbach's α) of the immediate and delayed recall posttests were both .91.

Table 5***Examples of Correct and Partially Correct Answers in Meaning Recall***

Correct answer types	Examples
Correct answers	Likely to hurt someone ¹
Partially correct answers (some but not all semantic features):	Violence; violent; hurt others
Partially correct answers (added semantic features):	Domestic violence; people who are likely to abuse others

1: This is also the L2 gloss given to participants.

Meaning Matching Tests

A response was deemed correct when it matched the target word with its corresponding meaning. Correct responses were coded as 1 and incorrect ones as 0. The total data points collected and analyzed for each participant in this test was 24. The immediate and delayed matching posttests had a reliability estimate (Cronbach's α) of .92 and .90 respectively.

Self-paced Reading Tests

Two types of data were collected from the self-paced reading test: accuracy data from the comprehension question at the end of each trial and RT data from sentence reading. The accuracy data from the comprehension questions were used for data cleaning. First, participants with an accuracy rate of lower than 70% were removed. This affected two participants in the immediate posttest and six in the delayed posttest. Subsequently, trials where participants did not correctly answer the comprehension questions were also removed.

RT data to four regions in the self-paced reading test were examined, namely the critical position (i.e., where the pseudowords, nonwords, or real words appeared, depending on the

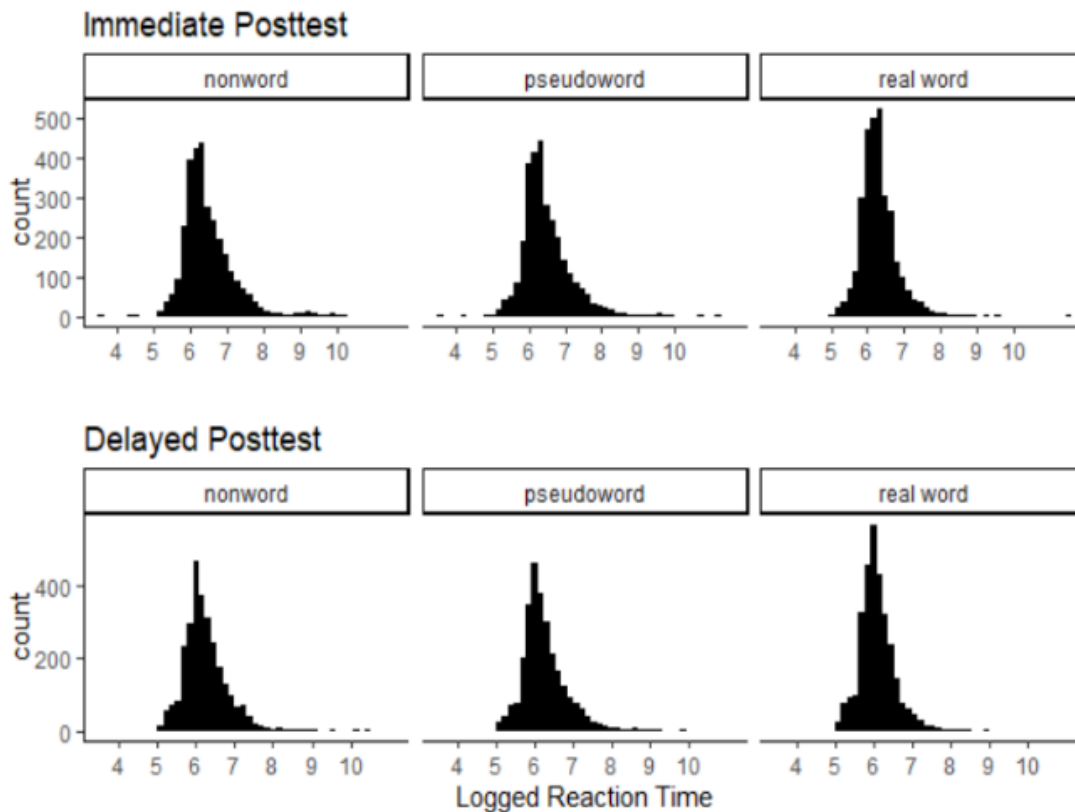
condition), the word before the critical position (henceforth position 0), and the two words that follow (henceforth position 2 and position 3). RT to position 0 was used for a manipulation check: no significant difference should be found in RTs at this position in the three conditions (i.e., pseudoword, nonword, and real word) to make sure that participants had the same starting point before reaching the critical position. The critical position, positions 2, and 3 were selected because difficulty of word meaning retrieval and integration may not manifest in RTs to the critical position but to the immediate word that follows, i.e., the spill-over effect; for L2 learners in particular, the spill-over effect may also occur at the second word after the critical position due to lower processing efficiency (Jiang, 2013). The critical RT comparisons were between the pseudoword and nonword conditions, and between the pseudoword and real word conditions. If learners were able to retrieve and integrate the meanings of the pseudowords in a reading task, the RTs to the critical position, position 2, or 3 in the pseudoword condition should be faster than those in the nonword condition, in which the learners would read a nonword they had never encountered. The comparison of RTs between the pseudoword and real word conditions would tell us whether the newly-learned pseudowords could be retrieved as fluently as familiar real words.

RT trimming was implemented using the *trimr* package (version 1.1.1; Grange, 2015) in R. Figure 8 shows the distribution of RTs in the immediate and delayed self-paced reading posttests on a natural log scale. The RT distributions in the three conditions were slightly different: RTs in the real word condition concentrated between 6 and 7 on the log scale while there were larger proportions of RTs greater than 7 in the nonword and pseudoword conditions. In this case, using absolute values alone may be inappropriate for cleaning the RT data. Therefore, I chose to use standard deviation, along with absolute values, to identify outliers.

Specifically, I chose 250 ms as the low cutoff. Responses with RTs below this value were removed because they were too fast to have accurately reflected the genuine reading process. In addition, responses with RTs 2.5 *SD* away from the mean RT of a condition were excluded. The method of identifying outliers based on *SD* took into account the different distributions of RTs in each condition. RT cleaning affected 4.06% and 7.53% of the data in the immediate and delayed self-paced tests respectively.

Figure 8

Reaction Time Distribution in the Immediate and Delayed Self-Paced Reading Posttests



Mixed-effects Modelling

To answer the research questions, I used mixed-effects modelling to analyze gloss time data recorded during participants' reading, accuracy data from the meaning matching and recall

posttests, and RT data from the self-paced reading test. In all mixed-effects models, I used treatment coding for the categorical variable of group (L1 and L2 gloss groups), with L1 gloss group as the reference. All continuous independent variables (i.e., target word FoO, vocabulary size, and gloss time) were standardized and mean-centered using the *scale()* function in R to reduce collinearity.

A maximal model was first built, which included (1) main effects of theoretical interest; (2) justifiable interactions between the main effects of interest (see The Present Study section); and (3) maximal random-effects structure justified by the data (Barr et al., 2013). For details of the initial maximal models for each analysis, see Appendix H. Two steps were involved in model selection (Gries, 2021). The first step was to determine the random-effects structure: when there was a convergence issue, the random-effects structure was simplified by dropping by-subject random slopes first (Barr et al., 2013), followed by by-item random slopes; then, I continued to simplify the random-effects structure by dropping elements that accounted for the least variability (i.e., those having the smallest standard deviation in the random effects table in R output) one by one. Akaike Information Criterion (AIC) was used to select models with the best random-effects structure, i.e., the model with the smallest AIC. After the random-effects structure was determined, the second step was to select the best fixed-effects structure. This step involved excluding insignificant interactions, followed by dropping interactions that did not improve model fit. Main effects of theoretical interests did not participate in model selection and were all kept. The best model in the second step would be the model with the smallest AIC. 95% Wald confidence intervals for estimates in the final models were calculated through the *confint.merMod()* function.

After model selection, the selected model went through model diagnostics. First, residuals of the final models were checked for outliers. Observations with a residual greater than 2.5 *SDs* away from the mean were removed, after which the final models were refitted. Second, VIFs as a test for multicollinearity of the final models were checked using the performance package (Lüdtke et al., 2020). A VIF value under 10 would indicate that there was no significant multicollinearity (Hair et al., 1995).

RQ1 looked into gloss language effect on learners' gloss engagement, and how the gloss language effect on gloss engagement was moderated by learners' vocabulary size. The gloss time data contained a large number of zeros ($n = 516$, representing 19% of the data). Data with such distribution, i.e., a portion of zeros, plus a continuous non-zero part, is called semicontinuous data and is common in biomedical and econometric research. Because zeros and non-zero data points are often viewed as results of two distinct processes, a two-part mixed model has been proposed to analyze such data (e.g., Olsen & Schafer, 2001; Tu & Zhou, 1999). The two-part mixed model fits zero and non-zero data separately. Zeros are treated as binary data (i.e., zeros vs. non-zeros) and are modelled with a generalized linear mixed model. Non-zero data are viewed as continuous and are analyzed with a linear mixed model. In the current study, zeros and non-zero data can be seen as representing two processes. The former represented a lack of processing on a gloss. The latter denoted the actual amount of processing on a gloss. Therefore, I followed the two-part mixed model and analyzed the data in two ways. The first analysis examined gloss language effect on whether participants processed a gloss or not, and how vocabulary size may moderate this gloss language effect. To do so, I created a binary variable called 'gloss checking', where I coded the zeros in the gloss time data as 0, representing 'no gloss processing', and the non-zeros as 1, representing 'gloss processing'. A generalized linear

mixed-effects model was built, with gloss processing as the dependent variable, and group (L1 vs. L2 gloss) and vocabulary size as the main effects of theoretical interest. In the second analysis, I removed zeros from the gloss time data. I then logged transformed the data to bring its distribution closer to normal and fit a linear mixed-effects model to the log transformed gloss time data that did not contain zero. Group (L1 vs. L2 gloss) and vocabulary size were the main effects of theoretical interest.

RQ2 examined the effect of gloss language on short-term and long-term learning, and the moderating effects of target word FoO, vocabulary size, and engagement on gloss language effect. To answer RQ2, accuracy data from the meaning matching and recall posttests, and RT data from the self-paced reading test were analyzed. First, four generalized linear mixed-effects models were built, each for accuracy data from the meaning matching immediate and delayed posttests, and meaning recall immediate and delayed posttest. In these models, correct responses were coded as 1 and incorrect ones as 0. The main effects of theoretical interest in these models were group (L1 vs. L2 gloss), FoO, vocabulary size, and time on gloss.

For the self-paced reading tests, data from each group (i.e., L1 and L2 gloss) and from each test timing (i.e., immediate and delayed) were analyzed separately. Data from each position (i.e., positions 0, critical & 2) were also analyzed in different models. In other words, separate analyses were conducted by group, position, and test timing. Based on the descriptive statistics of the RT data, RTs to position 3 were not analyzed. This was because the descriptive statistics indicated that the RT difference between conditions was not likely to be found at this position, i.e., the spill-over effect was not likely to have spread to this position. For each group and each test timing, a linear mixed-effects model was built for RTs each to the critical position and position 2 (i.e., the position following the critical position). The main effects of theoretical

interest in these mixed-effects models were condition (i.e., nonword, pseudoword, and real word), FoO, vocabulary size, and time on gloss. In addition, a manipulation check was conducted to see if the RT difference between conditions at position 0 was significant. The main effect of interest was condition in the mixed-effects models for the manipulation check. In all models for the self-paced reading tests, RTs were log transformed to bring the distribution closer to normal. In addition, condition was treatment coded, with the pseudoword condition being the baseline. This allowed the crucial comparisons between the real word and pseudoword conditions, and between the nonword and the pseudoword conditions to be presented and interpreted in a more straightforward manner.

CHAPTER 3: RESULTS

This chapter first reports descriptive statistics of the dependent variables, namely time on gloss, number of correct answers in the meaning matching and meaning recall posttests (immediate and delayed), and reaction times (RTs) to the four regions of interests in the immediate and delayed self-paced reading tests. Next, the chapter presents the final mixed-effects models from the analyses engaged to answer the research questions.

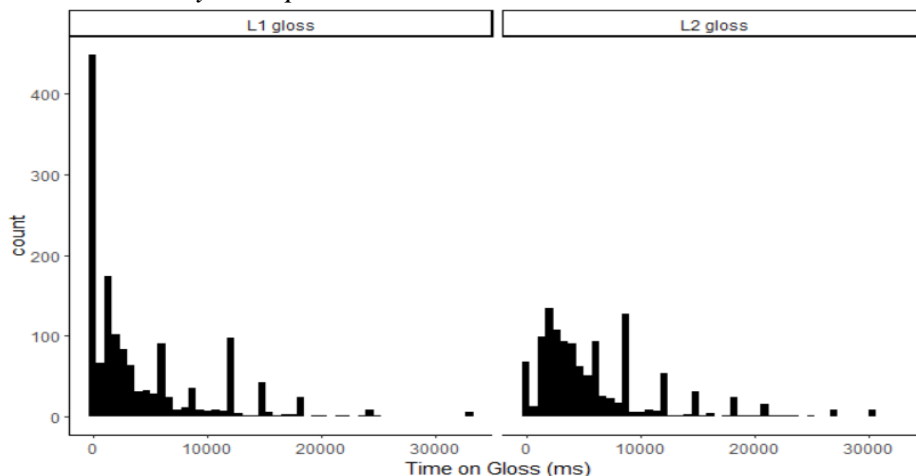
Descriptive Statistics

Time on Gloss

After gloss time cleaning (see Time on Gloss in the Data Analysis section), the average time a participant spent on a gloss was 4171 milliseconds (ms) ($SD = 5426.67$; range: 0 – 33000) in the first language (L1) gloss group and 6069 ms ($SD = 5472.83$; range: 0 – 30000) in the second language (L2) gloss group. Figure 9 plots the time on gloss by group. The histogram shows that the L1 gloss group had a lot more zeros, indicating that fewer participants spent time on the glosses in the L1 gloss group than in the L2 gloss group.

Figure 9

Time on Gloss by Group



Meaning Recall Tests

Table 6 shows the total number of correct responses (maximum = 24) in the meaning recall immediate and delayed posttests. Participants in both groups correctly recalled the meanings of less than 50% of the target words. The L1 gloss group did better than the L2 gloss group in both the immediate and delayed posttests.

Table 6

Total Number of Correct Responses in Meaning Recall Posttests by Group

	Immediate posttest		Delayed posttest	
	Mean (<i>SD</i>)	95% CI	Mean (<i>SD</i>)	95% CI
L1 Gloss	9.67 (6.68)	[7.94, 11.39]	8.10 (6.59)	[6.39, 9.80]
L2 Gloss	7.84 (5.20)	[6.36, 9.32]	6.30 (4.75)	[4.95, 7.65]

Meaning Matching Tests

Table 7 presents the total number of correct responses (maximum = 24) in the meaning matching immediate and delayed posttests. Both groups did better in receptive meaning recognition as measured by the meaning matching tests than in productive meaning recall as measured by the recall tests. The L1 gloss group correctly recognized around half (51.96%) of the target words in the immediate posttest but less than half (33.75%) in the delayed posttest. The L2 gloss group recognized less than half in both the immediate (42.33%) and delayed (37.67%) posttests. As in the meaning recall tests, the L1 gloss group performed better than the L2 gloss group in both the immediate and delayed posttests.

Table 7*Total Number of Correct Responses in Meaning Matching Posttests by Group*

	Immediate posttest		Delayed posttest	
	Mean (<i>SD</i>)	95% CI	Mean (<i>SD</i>)	95% CI
L1 Gloss	12.47 (6.68)	[10.74, 14.19]	10.38 (6.28)	[8.75, 12.02]
L2 Gloss	10.16 (6.18)	[8.41, 11.91]	9.04 (5.38)	[7.50, 10.58]

Self-paced Reading Test

The total number of observations included in the self-paced reading tests analyses was 8612 in the immediate posttest and 8000 in the delayed posttest. Table 8 and Table 9 summarize the means and standard deviations (in parenthesis) of RTs in the immediate and delayed posttests respectively. Figure 10 presents RTs of each position, and Figure 11 focuses on RTs to the critical position and position 2. In both the immediate and delayed posttests, positions 0 and 3 saw no large differences in RTs among the three conditions. That is, the participants started at a similar place before entering the critical position, where a nonword, a pseudoword, or a real word was inserted. In addition, the spill-over effect seemed to be at position 2, instead of at position 3. Focusing on the critical position, in both the immediate and delayed posttests, participants' RTs to nonwords and pseudowords were much larger than those to the real words, indicating processing difficulties when reading unfamiliar or newly learned words; the difference in RTs to nonwords and pseudowords was small at this position. For position 2, where a spill-over effect will likely happen, i.e., where processing difficulty is reflected, different RT patterns can be found for the L1 gloss and L2 gloss groups in the immediate posttest. For the L1 gloss group, RTs to the pseudowords seemed to be much higher than those to the nonwords. It seems that participants recovered faster from processing difficulties after reading a nonword than after

reading a pseudoword at the critical position. For the L2 gloss group, RTs to pseudowords and nonwords were similar. In the delayed posttests, the L1 gloss group's RTs at position 2 were slightly faster in the pseudoword than in the nonword conditions, while the L2 gloss groups read the stimuli in both conditions at a similar speed. RTs in the real word condition were much lower than those in the pseudoword and nonword conditions in both groups in the immediate and delayed posttests. RTs in the delayed posttest were lower in general than those in the immediate posttest, indicating a degree of familiarity with the stimuli when participants took the test for a second time.

Table 8

Reaction Time (ms) by Position and Condition in Self-paced Reading Immediate Posttest

	L1 gloss			L2 gloss		
	Real word	Pseudoword	Nonword	Real word	Pseudoword	Nonword
Position 0	536.67 (226.35)	573.53 (379.08)	530.13 (252.22)	603.44 (402.36)	595.15 (349.47)	586.16 (281.49)
Critical position	720.08 (513.11)	1194.23 (949.83)	1155.38 (861.04)	783.90 (546.55)	1230.91 (852.81)	1263.56 (844.87)
Position 2	614.27 (386.34)	797.46 (700.50)	685.72 (395.40)	671.83 (455.79)	799.66 (523.25)	778.96 (461.31)
Position 3	585.54 (431.37)	596.23 (429.66)	584.66 (321.19)	606.38 (329.01)	663.41 (417.57)	634.38 (370)

Table 9*Reaction Time (ms) by Position and Condition in Self-paced Reading Delayed Posttest*

	L1 gloss			L2 gloss		
	Real	Pseudoword	Nonword	Real	Pseudoword	Nonword
	word			word		
Position 0	454.86	482.15	474.87	490.51	496.19	470.01
	(194.56)	(275.26)	(244.25)	(234.91)	(267.76)	(190.64)
Critical	549.20	820.58	818.27	621.48	936.25	892.61
position	(296.34)	(667.65)	(717.82)	(334.93)	(793.12)	(661.65)
Position 2	505.53	635.09	697.29	529.96	685.55	701.12
	(252.91)	(398.60)	(532.31)	(251.56)	(343.02)	(443.33)
Position 3	472.16	507.98	483.73	506.52	554.95	557.59
	(219.27)	(238.55)	(178.96)	(261.34)	(301.98)	(326.89)

Figure 10

RTs in Self-paced Reading Tests by Group, Position, and Condition

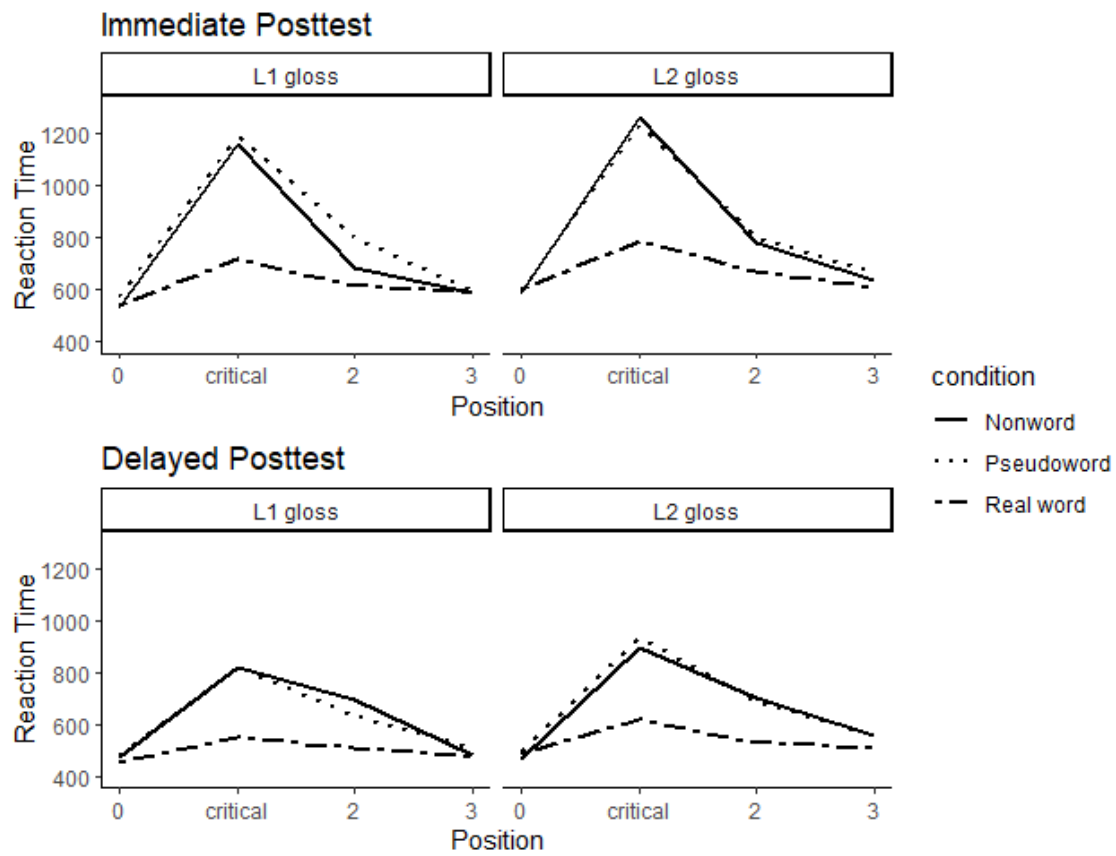
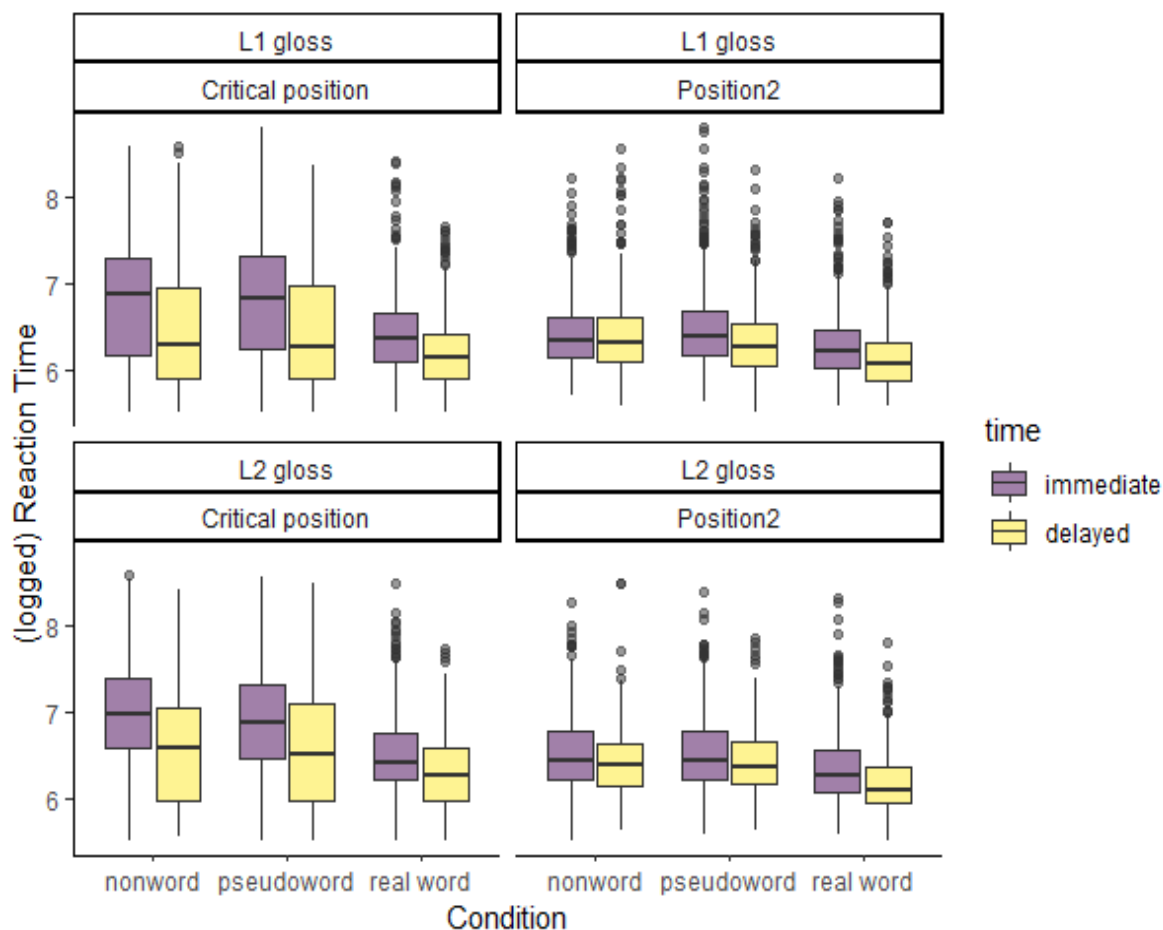


Figure 11
RT of the Critical Position and Position 2



RQ1

RQs 1a and b examined gloss language effect on gloss engagement and the potential moderating role of vocabulary size on the gloss language effect. I first investigated whether gloss language influenced the likelihood of participants processing a gloss. Table 10 showed that the L2 gloss group were significantly more likely than the L1 gloss group to process a gloss. While vocabulary size did not moderate the gloss language effect, it predicted gloss processing: participants with a larger vocabulary size were more likely to spend time processing a gloss.

The analysis on gloss time showed a different pattern (see Table 11). Neither Group nor vocabulary size were significant. The L1 gloss and L2 gloss groups spent similar amount of time on the glosses, i.e., no gloss language effect. Vocabulary size did not have an effect on gloss time.

Table 10

Mixed-effects Model for Gloss Engagement: zero vs. non-zero gloss time data

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>z</i>	<i>P</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	1.52 [.69, 2.35]	.42	3.60	<.001***	5.67	2.38	1.64	1.28
Group (L2 gloss)	3.80 [2.49, 5.11]	.67	5.68	<.001***			1.88	1.37
Vocabulary size	.59 [.08, 1.10]	.26	2.27	.02*				

Table 11*Mixed-effects Model for Gloss Engagement: Time on Gloss*

	Fixed effects				Random effects			
					By participant		By item	
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	8.24 [8.02, 8.45]	.11	76.05	<.001***	.26	.51	.16	.42
Group (L2 gloss)	.22 [-.04, .48]	.13	1.64	.11			.32	.57
Vocabulary size	-.02 [-.12, .08]	.05	-.34	.74				

RQ2

RQs 2a and b asked how gloss language affected learning and how target word frequency of occurrence (FoO), learners' vocabulary size, and learner engagement moderated the effects of gloss language on learning. Learning was measured by three tests, namely meaning matching, meaning recall, and self-paced reading tests, at two time points, i.e., immediate and delayed. In the following sections, I present findings by test type and test time.

Meaning Recall Test

Table 12 and Table 13 present the final mixed-effects model for accuracy data in the meaning recall immediate and delayed posttests respectively. For RQ2a, in both the immediate and delayed posttests, the main effect of group was not significant, indicating that when all other variables were at their mean standardized values, the two gloss language groups did not differ

significantly in their performance in the meaning recall posttests. In both the immediate and delayed posttests, target word FoO and participants' time on reading the glosses were significant predictors of learning while the main effect of vocabulary size was not significant.

Table 12
Meaning Recall Immediate Posttest Mixed-effects Model

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>z</i>	<i>p</i>	By participant		By item	
					<i>Varianc</i>	<i>S</i>	<i>Varianc</i>	<i>S</i>
					<i>e</i>	<i>D</i>	<i>e</i>	<i>D</i>
Intercept	-.81[-1.48, -0.14]	.34	-2.38	.02*	4.29	2.07	.85	.92
Group (L2 gloss)	-.50 [-1.32, .33]	.42	-1.18	.24				
FoO	1.32 [.93, 1.72]	.20	6.61	<.001***				
Vocabulary size	.30 [-.11, .72]	.21	1.42	.15				
Time on gloss	.33 [.12, .55]	.11	3.08	.002**				
Group * Time on gloss	-.36 [-.63, -.07]	.14	-2.51	.01*				

Table 13
Meaning Recall Delayed Posttest Mixed-effects Model

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>z</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	-1.18 [-1.75, -.61]	.29	-	<.001***	3.07	1.75	.60	.77
Group (L2 gloss)	-.51 [-1.22, .20]	.36	-	.16				
FoO	.90 [.55, 1.25]	.18	5.02	<.001***				
Vocabulary size	.26 [-.23, .75]	.25	1.03	.30				
Time on gloss	.39 [.19, .60]	.10	3.76	<.001***				
Group * FoO	.07 [-.17, .30]	.12	.55	.58				
Group * Vocabulary size	-.48 [-1.20, .23]	.36	-	.19				
FoO * Vocabulary size	.26 [.09, .42]	.09	3.00	.003**				

Table 13 (cont'd)

Group * Time	-.35 [-.63, .14	-	.01*
on gloss	-.08]	2.50	
FoO * Time on	-.24	.12	-.05
gloss	[-.48, .003]	1.93	
Group * FoO *	-.19	.12	-.10
Vocabulary	[-.43, .04]	1.64	
size			
Group * FoO *	.42 [.09, .75]	.17	2.47 .01*
Time on gloss			

For RQ2b, several interactions were found. First, in the immediate recall test, time on gloss significantly moderated the effect of gloss language. Specifically, the more time participants spent on a gloss, the larger the difference between the L1 gloss and L2 gloss groups was, due largely to increase in the accuracy of the L1 gloss group (see Figure 12). Second, in the delayed posttest, the effect of gloss language was moderated by both time on gloss and target word FoO. Figure 13 plots the three-way interaction between group, time on gloss, and FoO. Figure 13a shows the gloss time effect for each FoO range by group. Figure 13b juxtaposes group performance as affected by gloss time by FoO range. Looking at Figure 13a, for the L1 gloss group, the positive effect of time on gloss was the largest when FoO was low. The positive effect of time on gloss became smaller when FoO increased and even became negative when FoO reached its largest values. In contrast, for the L2 gloss group, the effect of time on gloss first increased as FoO became larger. The gloss time effect became the largest when FoO was at mid-

range (z score: 0.5 – 1.5; raw FoO: 10 – 13). The effect of gloss time then decreased as FoO increased, but remained positive. Figure 13b indicated that the gloss language effect varied by time on gloss and FoO. When FoO was relatively low (z score: -1.5 – -0.5; -0.5 – 0.5; raw FoO: 1 – 3; 4 – 9), the L1 gloss group outperformed the L2 gloss group, and this gloss language effect became larger as time on gloss increased. When FoO was higher (z score: 0.5 – 1.5; 1.5 – 2.5; 2.5 – 3.5; raw FoO range: 10 – 13; 14 – 19; 20 – 27), the L2 gloss group seemed to be able to outperform the L1 gloss group when participants spent more time on the glosses. It seems that in the delayed posttest, the L2 gloss group may recall the meanings of more words if participants had sufficient encounters with the target words and sufficient reading time on the L2 glosses. The three-way interaction, however, should be interpreted with caution due to the small number of target words with a raw FoO above 10 ($n = 3$).

Figure 12

*Group * Time on Gloss Interaction (Meaning Recall Immediate Posttest)*

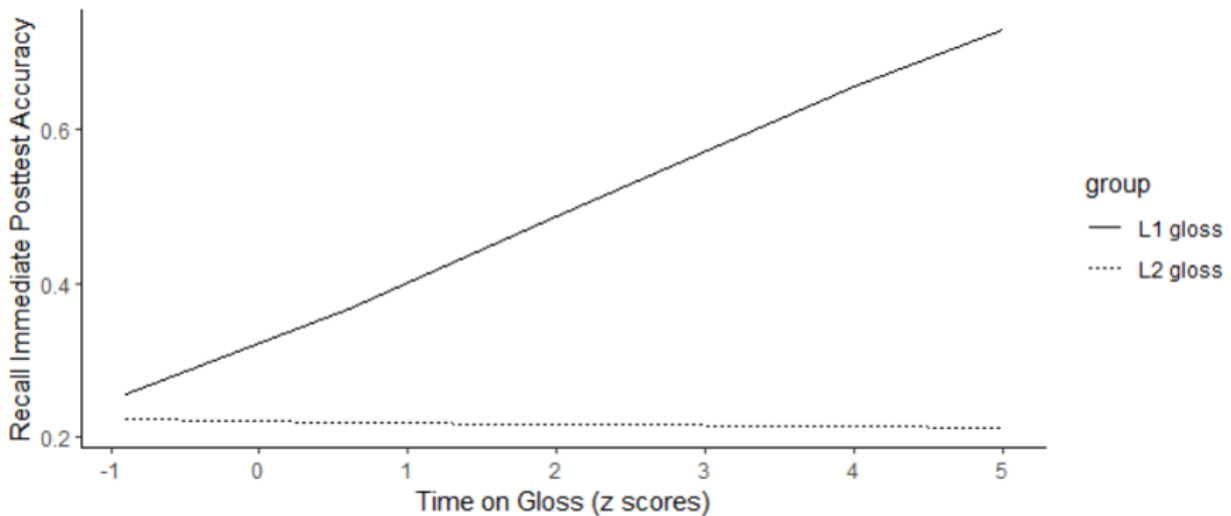
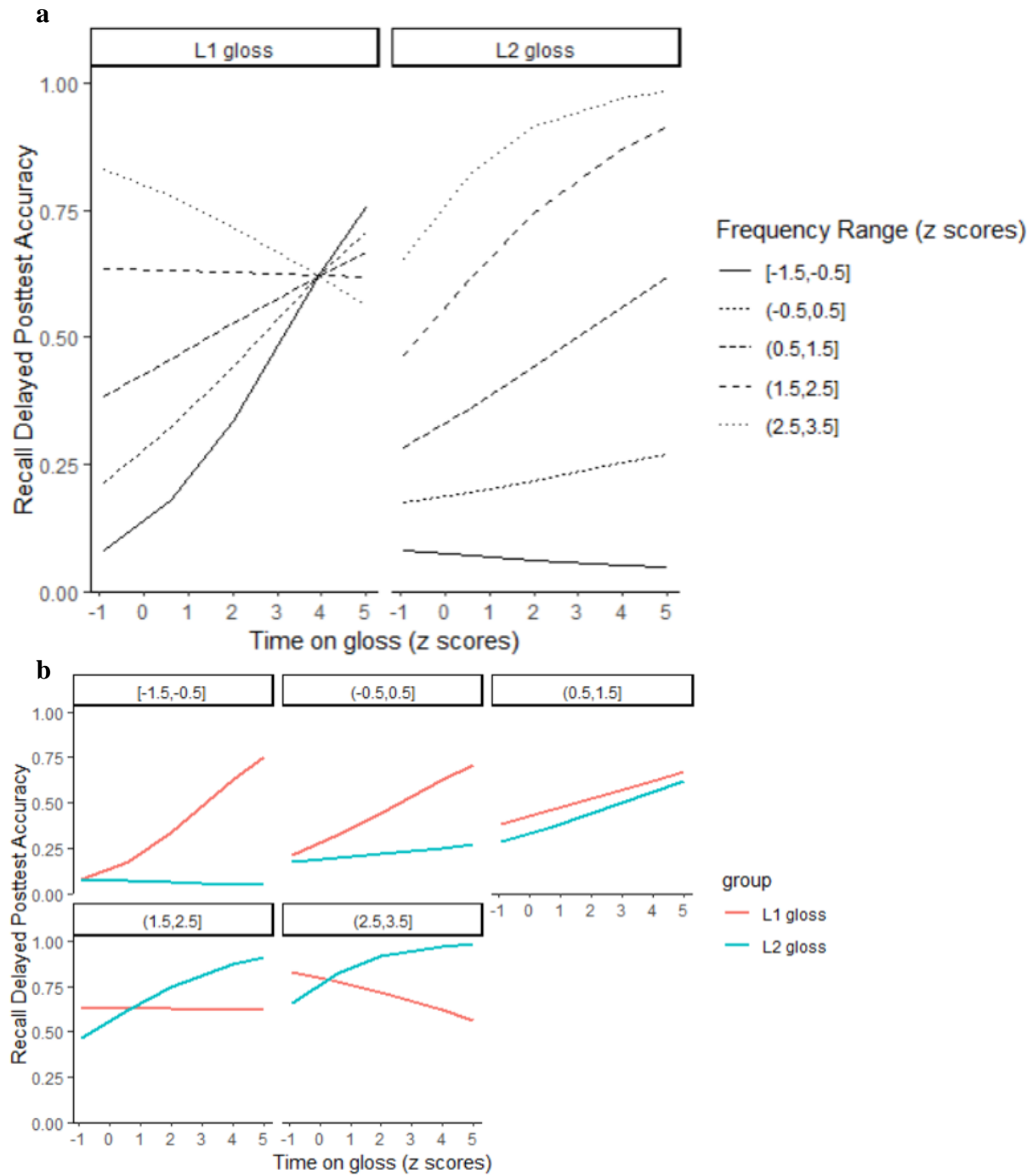


Figure 13

*Group * Time on gloss * FoO Interaction (Meaning Recall Delayed Posttest)*



Meaning Matching Test

The final mixed-effects models for the meaning matching immediate and delayed posttests were presented in Table 14 and Table 15. For RQ2a, in both posttests, results showed no significant difference between the L1 gloss and L2 gloss groups, i.e., no significant gloss language effect, when other variables were held at their mean standardized values. FoO, vocabulary size, and the amount of time spent on the glosses all had positive effects on learning. For RQ2b, in the immediate posttest, the variable of group did not interact with any other variables. In the delayed posttest, there was a borderline significance for the interaction between group and time on gloss (RQ2b). As Figure 14 shows, the more time the L1 gloss group spent on glosses, the higher the accuracy was, leading to larger group difference, i.e., larger gloss language effect. The group and time on gloss interaction pattern in the meaning matching delayed posttest was similar to that in the meaning recall immediate posttest.

Table 14
Meaning Matching Immediate Posttest Mixed-effects Model

	Fixed effects					Random effects			
						By participant		By item	
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>z</i>	<i>p</i>		<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	.09 [-.51, .68]	.30	.29	.78		3.41	1.85	.69	.83
Group (L2 gloss)	-.63 [- 1.36, .11]	.37	-	.09					
FoO	.98 [.63, 1.33]	.18	5.45	<.001***					
Vocabulary size	.57 [.07, 1.08]	.26	2.21	.03*					
Time on gloss	.22 [.08, .35]	.07	3.04	.002*					
Group *	-.71 [-	.38	-	.06					
Vocabulary size	1.45, .02]		1.90						

Table 15
Meaning Matching Delayed Posttest Mixed-effects Model

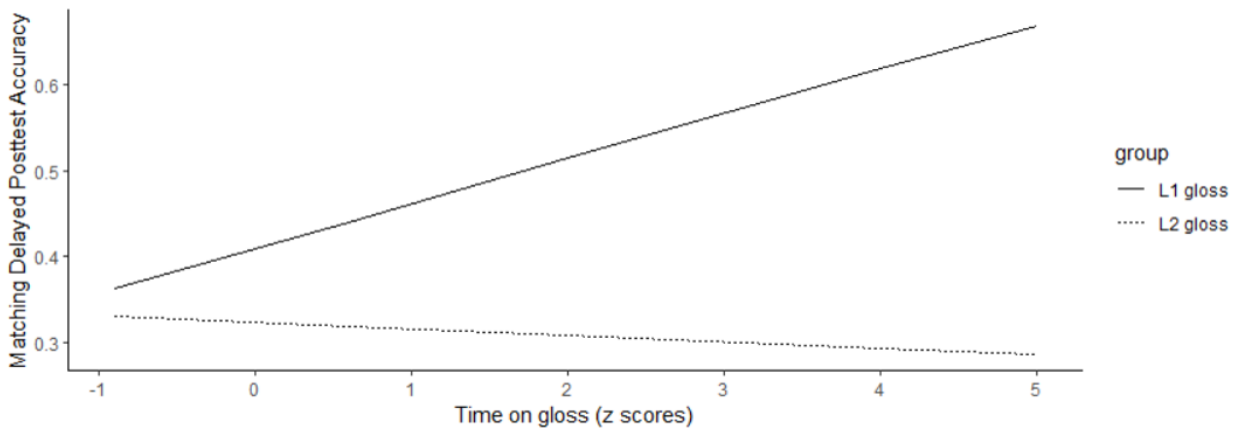
	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>z</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	-.38 [-.87, .11]	.25	-	.13	2.27	1.51	.43	.66
Group (L2 gloss)	-.39 [- 1.00, .22]	.31	-	.21				
FoO	.86 [.55, 1.16]	.16	5.48	<.001***				
Vocabulary size	.43 [.01, .85]	.21	2.02	.04*				
Time on gloss	.21 [.02, .39]	.09	2.15	.03*				
Group * FoO	-.02 [-.24, .21]	.12	-.14	.88				
Group * Vocabulary size	-.58 [- 1.20, .03]	.31	-	.06				
FoO * Vocabulary size	.19 [.03, .35]	.08	2.37	.02*				

Table 15 (cont'd)

Group * Time	-.26 [-.51, .13	-	.02*
on gloss	-.01]	2.02	
FoO * Time on	-.27 [-.51, .12	-	.02*
gloss	-.05]	2.37	
Group * FoO *	-.08	.11	-.69 .49
Vocabulary	[-.29, .14]		
size			
Group * FoO *	.29 [-.03, .60]	.16	1.80 .07
Time on gloss			

Figure 14

*Group * Time on Gloss Interaction (Meaning Matching Delayed Posttest)*



Self-paced Reading Test

The manipulation check indicated that RTs for position 0 did not differ significantly between the pseudoword and the nonword conditions, and between the pseudoword and real

word conditions for the L1 and L2 gloss groups in both the immediate and delayed posttests (see details of the analyses in Appendix H). The results of the manipulation check suggested that participants started at roughly the same place before encountering the pseudoword, nonword, or real word at the critical position. In other words, effects found at the critical position and/or at position 2 may be attributed not to pre-existing processing differences at position 0, but to differences in the processing of the nonwords, pseudowords, and real words at the critical position. In what follows, I present results of the immediate posttest first, followed by those of the delayed posttest. For each test time point, I present the results of the L1 gloss group, including results of the critical position and position 2, which is then followed by the results of the L2 gloss group.

Immediate Posttest, L1 Gloss. Table 16 and Table 17 include results for the critical position and position 2 respectively for the L1 gloss group in the immediate posttest. At the critical position, RTs for the pseudowords did not differ significantly from those for nonwords, while the pseudoword RTs were significantly slower than those for the real words. The comparison of RTs for these three types of stimuli reflected participants' processing difficulties of nonwords and pseudowords. There was a significant interaction between condition and time on gloss, showing that the more time participants spent on a gloss, the larger the RT difference was between the pseudowords and the real words, owing primarily to the speeding up in reacting to the real words (see Figure 16).

At position 2, RTs for the pseudoword condition were significantly higher than those for both the nonword and the real word conditions. However, when participants encountered a pseudoword more times in the reading (i.e., higher target word FoO), RTs for the pseudoword conditions decreased (i.e., significant main effect of FoO). In turn, the RT difference between the

pseudoword and real word conditions decreased (i.e., significant interaction between condition and FoO). When FoO reached the highest (i.e., 27 times), the RTs for the real word and to the pseudoword conditions were close (see Figure 15). Time on gloss also interacted with condition, albeit in a different direction. The more time participants spent on a gloss, the larger the RT difference was between the pseudoword and the real word conditions, and between the pseudoword and the nonword conditions, mostly due to the decrease in RTs for the real word condition and increase in RTs for the nonword condition (see Figure 17).

Table 16*Self-paced reading immediate posttest: Critical Position, L1 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.83 [6.71, 6.95]	.06	111.81	<.001***	.15	.38	.01	.11
Nonword	.01 [-.06, .08]	.04	.32	.75				
Real word	-.41[-.48, -.34]	.04	-11.69	<.001***				
FoO	-.003 [-.06, .04]	.03	-.12	.90				
Vocabulary size	-.04 [-.14, .06]	.05	-.75	.46				
Time on gloss	.04 [-.02, .10]	.03	1.20	.23				
Nonword * Time on gloss	-.004 [-.08, .07]	.04	-.09	.92				
Real word * Time on gloss	-.08 [-.15, -.01]	.04	-2.26	.02*				

Table 17*Self-paced reading immediate posttest: Position 2, L1 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.51 [6.44, 6.58]	.04	177.30	<.001 ***	.03	.18	.01	.09
Nonword	-.08 [-.13, -.02]	.03	-2.58	.01*				
Real word	-.21 [-.26, -.15]	.03	-7.24	<.001***				
FoO	-.08 [-.14, -.03]	.03	-2.91	.01*				
Vocabulary size	-.02 [-.07, .04]	.03	-.59	.56				
Time on gloss	.05 [.002, .10]	.02	2.02	.04*				
Nonword *	.07 [.01, .12]	.03	2.32	.02*				
FoO								
Real word *	.06 [.003, .12]	.03	2.07	.04*				
FoO								

Table 17 (cont'd)

Nonword * -.05 [-.11, .01] .03 -1.79 .07

Time on

gloss

Real word * -.07 [-.12, .03 -2.25 .02*

Time on -.01]

gloss

Figure 15

*Condition * FoO Interaction (Position 2, Immediate Self-paced Reading Test, L1 Gloss Group)*

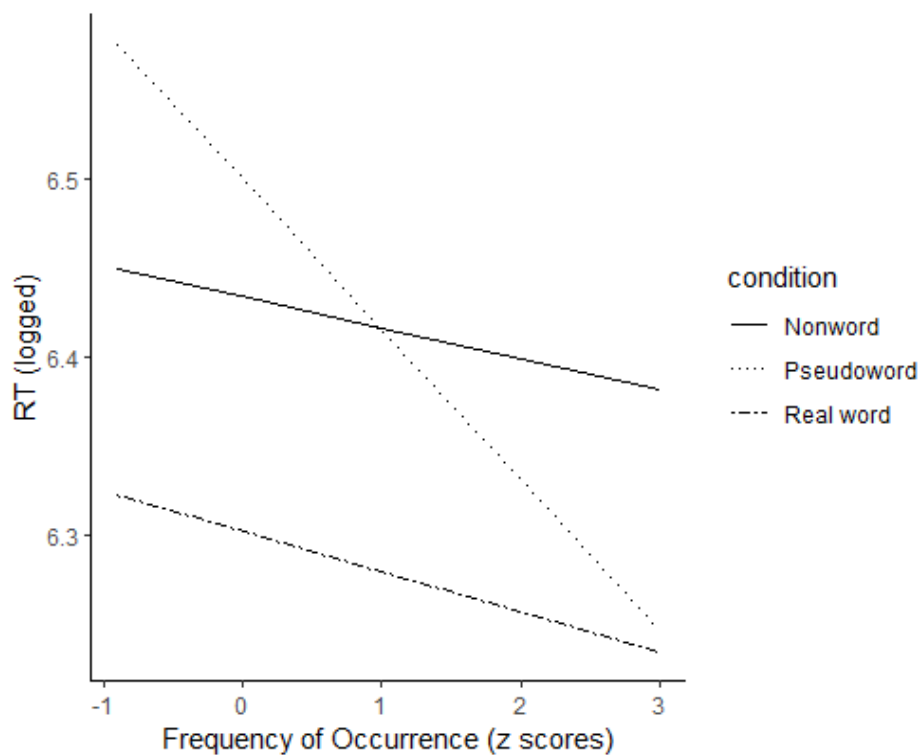


Figure 16

*Time on Gloss * Condition Interaction (Critical Position, Immediate Self-paced Reading Test, L1 Gloss Group)*

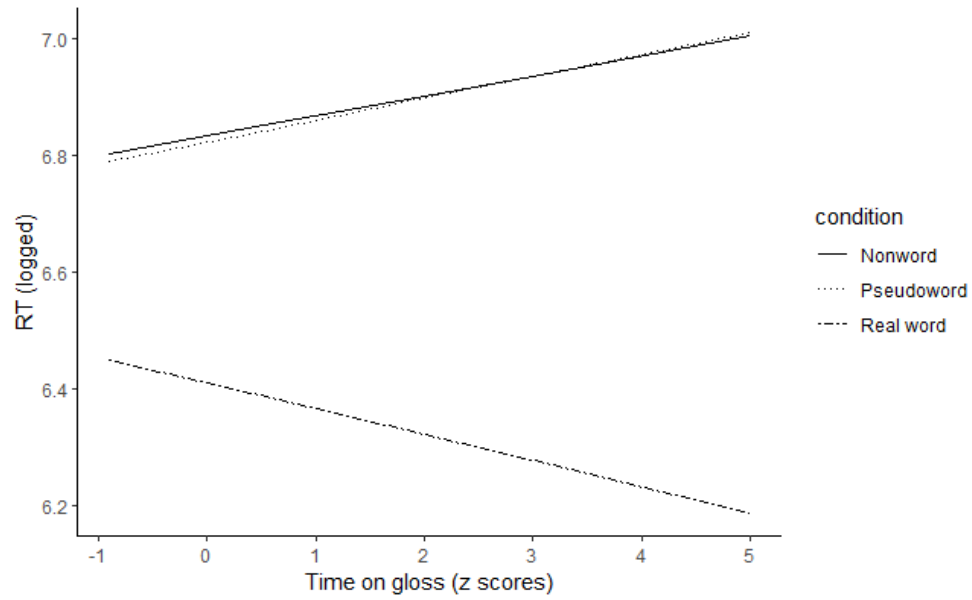
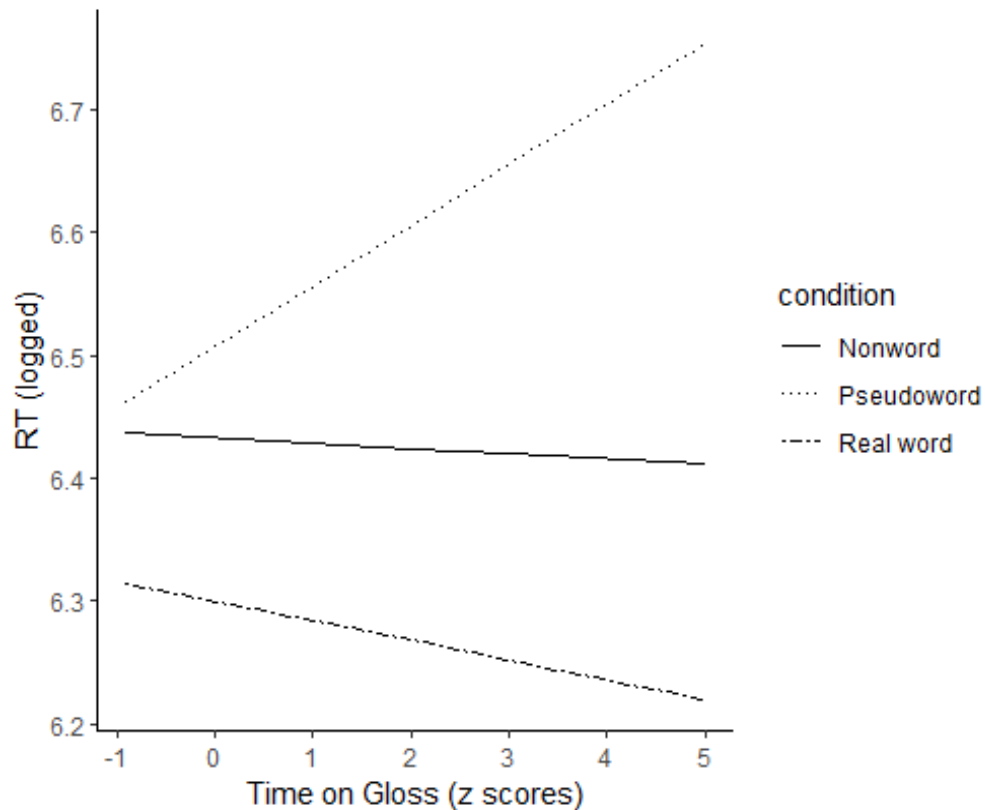


Figure 17

*Condition * Time on Gloss Interaction (Position 2, Immediate Self-paced Reading Test, L1 Gloss Group)*



Immediate Posttest, L2 Gloss. Results for the critical position and position 2 are presented in Table 18 and Table 19 respectively. Like the L1 gloss group, participants in the L2 gloss group reacted significantly slower to the pseudowords than to the real words and reacted similarly to the nonwords at the critical position, reflecting processing difficulties in reading nonwords and pseudowords. Unlike the L1 gloss group, RTs of the L2 gloss group to the three types of stimuli were not moderated by any other variables. RTs at position 2 showed a similar pattern to those at the critical position. That is, RTs in the pseudoword condition were similar to those in the nonword condition and were significantly higher than those in the real word condition. No other variables or interactions were found significant.

Table 18*Self-paced reading immediate posttest: Critical Position, L2 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.91 [6.77, 7.04]	.07	105.06	<.001***	.15	.38	.01	.11
Nonword	.01 [-.06, .09]	.04	.39	.70				
Real word	-.42 [-.49, -.34]	.04	-11.27	<.001***				
FoO	-.02[-.07, .04]	.03	-.63	.53				
Vocabulary size	.03[-.08, .14]	.06	.52	.61				
Time on gloss	.01[-.03, .04]	.02	.26	.80				

Table 19*Self-paced reading immediate posttest: Position 2, L2 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>P</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.52 [6.43, 6.61]	.05	142.29	<.001***	.06	.24	.01	.09
Nonword	.003 [-.06, .06]	.03	.09	.93				
Real word	-.16 [-.22, -.10]	.03	-5.15	<.001***				
FoO	-.04 [-.08, .01]	.02	-1.53	.14				
Vocabulary size	-.01 [-.08, .06]	.04	-.31	.76				
Time on gloss	.02 [-.01, .05]	.02	1.15	.25				

Delayed Posttest, L1 Gloss. RTs at the critical position (see Table 20) showed similar patterns to those in the immediate posttest: significantly faster response to real words than to pseudowords, and similar RTs to nonwords and pseudowords. A significant three-way interaction was found for condition, FoO, and time on gloss. Figure 18 plots this interaction, with each panel representing an FoO range. The interaction indicated that when FoO was low (range in z scores: -1.5 – -0.5; range in raw FoO: 1 – 3), the more time participants spent on a gloss, the slower they responded to the nonwords and pseudowords. As FoO increased, the relationship between time on gloss and RTs changed: more time on a gloss led to faster response to

pseudowords. When FoO was between -0.5 and 0.5 (raw FoO: 4 – 9), RTs to pseudowords got closer to those to real words as time on gloss increased. As FoO continued to increase (above z-score FoO of 0.5 and raw FoO of 9), sufficient time spent reading glosses may facilitate the processing of pseudowords, which eventually became faster than that of the real words. Again, because there were only three target words with a raw FoO above 10, this three-way interaction should be interpreted cautiously.

Table 20*Self-paced reading delayed posttest: Critical Position, L1 Gloss Group*

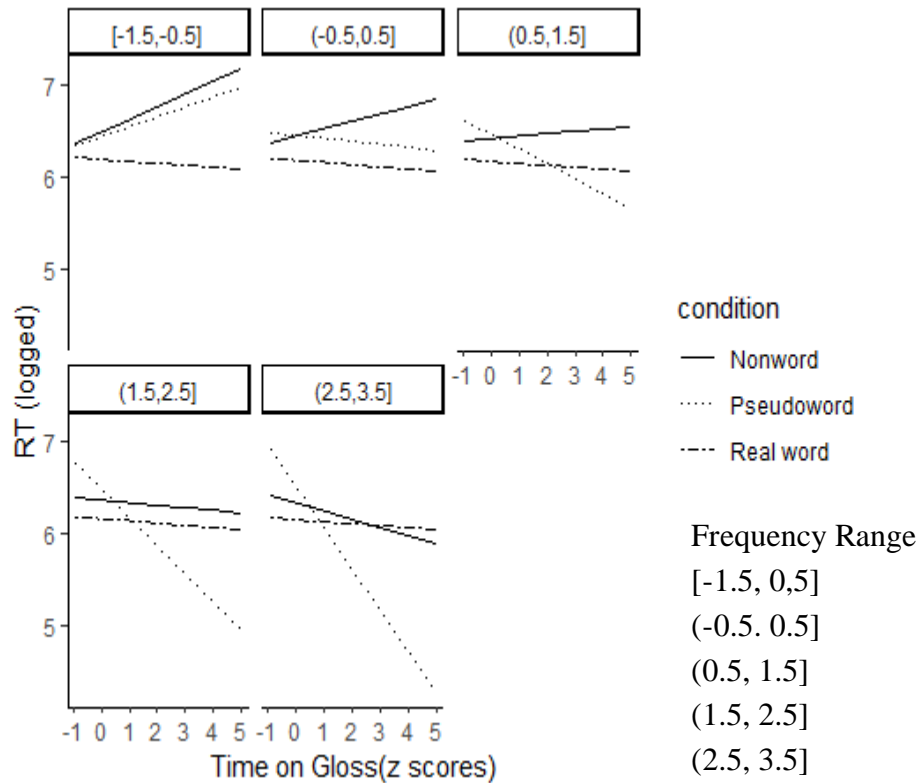
	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>T</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.45 [6.34, 6.57]	.06	113.50	<.001***	.13	.36	.01	.08
Nonword	-.004 [-.07, .07]	.04	-.11	.91				
Real word	-.27 [-.34, -.20]	.04	-7.69	<.001***				
FoO	.02 [-.05, .08]	.03	.53	.60				
Vocabulary size	-.03 [-.13, .07]	.05	-.64	.52				
Time on gloss	-.02 [-.08, .04]	.03	-.68	.50				
Nonword * FoO	-.06 [-.13, .02]	.04	-1.43	.15				
Real word * FoO	-.03 [-.10, .04]	.04	-.76	.45				
Nonword * Time on gloss	.10 [.03, .18]	.04	2.88	.004*				

Table 20 (cont'd)

Real word *	-.001	.04	-.04	.97
Time on	[-.07, .07]			
gloss				
FoO * Time	-.14 [-.23,	.04	-3.27	.001**
on Gloss	-.06]			
Nonword *	.09 [-.03, .20]	.06	1.46	.14
FoO * Time				
on gloss				
Real word *	.14 [.03, .26]	.06	2.44	.01*
FoO * Time				
on gloss				

Figure 18

*Condition * FoO * Time on Gloss Interaction (Critical Position, Delayed Self-paced Reading Test, L1 Gloss Group)*



At position 2, no significant difference was found in RTs between the nonword and the pseudoword conditions. Again, the RT difference was significant between the pseudoword and real word conditions. No variables moderated the RT difference among the three conditions (see Table 21 for results).

Table 21*Self-paced reading delayed posttest: Position 2, L1 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.31 [6.26, 6.40]	.04	176.99	<.001***	.04	.20	.003	.06
Nonword	.06 [.001, .111]	.03	1.99	.05				
Real word	-.19 [-.25, -.14]	.03	-6.90	<.001***				
FoO	-.02 [-.05, .02]	.02	-1.01	.33				
Vocabulary size	-.02 [-.08, .04]	.03	-.54	.59				
Time on gloss	-.006 [-.05, .04]	.02	-.26	.79				
Nonword *	-.05 [-.11, .0004]	.03	-1.95	.05				
Time on gloss								
Real word *	.04 [-.01, .10]	.03	1.46	.15				
Time on gloss								

Delayed Posttest, L2 Gloss. Like the L1 gloss group, at the critical position, the L2 gloss group reacted significantly faster to real words than to pseudowords, but the pseudoword RTs did not differ significantly from nonwords RTs (see Table 22). This RT pattern was moderated by participants' vocabulary size. Specially, the larger participants' vocabulary size was, the faster they reacted to nonwords, eventually making RTs for nonwords shorter than those for pseudowords. Vocabulary size did not have a significant effect on RTs for pseudowords (i.e., no main effect). Figure 19 plots the Vocabulary Size * Condition interaction.

Table 22*Self-paced reading delayed posttest: Critical Position, L2 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.57 [6.45, 6.70]	.06	101.42	<.001***	.15	.38	.004	.06
Nonword	-.02 [-.10, .06]	.04	-.45	.65				
Real word	-.27 [-.35, -.19]	.04	-6.93	<.001***				
FoO	.002 [-.04, .04]	.02	.10	.92				
Vocabulary size	.05 [-.07, .17]	.06	.77	.44				
Time on gloss	.03 [-.01, .07]	.02	1.41	.16				
Nonword *	-.10 [-.17, -.02]	.04	-2.47	.01*				
Vocabulary size								
Real word *	-.07 [-.14, .01]	.04	-1.82	.07				
Vocabulary size								

Figure 19

*Condition * Vocabulary Size Interaction (Critical Position, Delayed Self-paced Reading Test, L2 Gloss Group)*

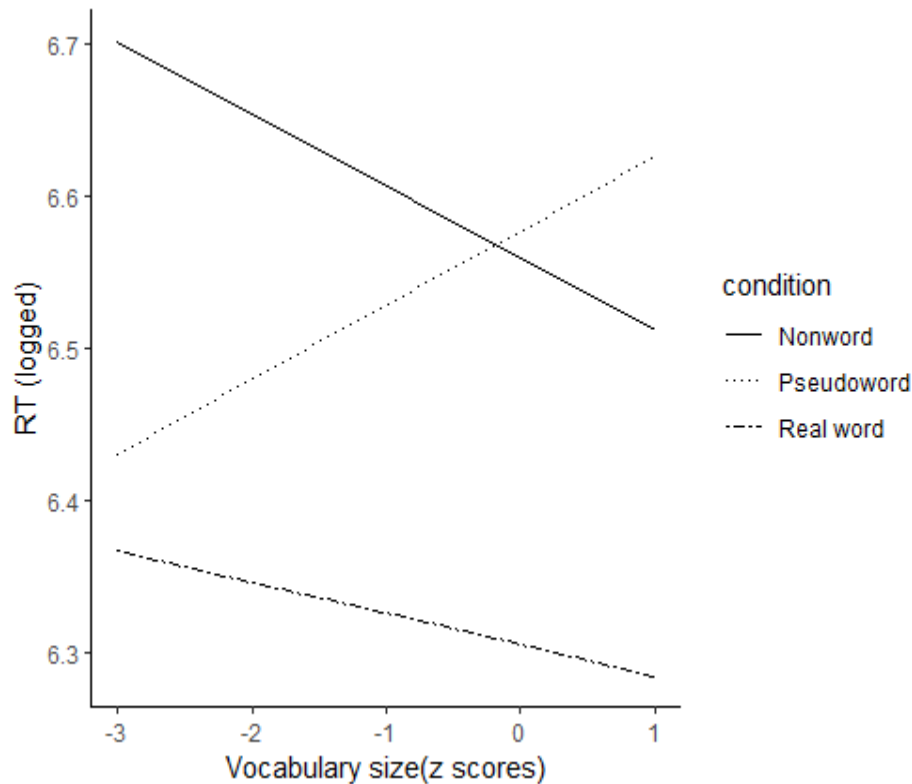


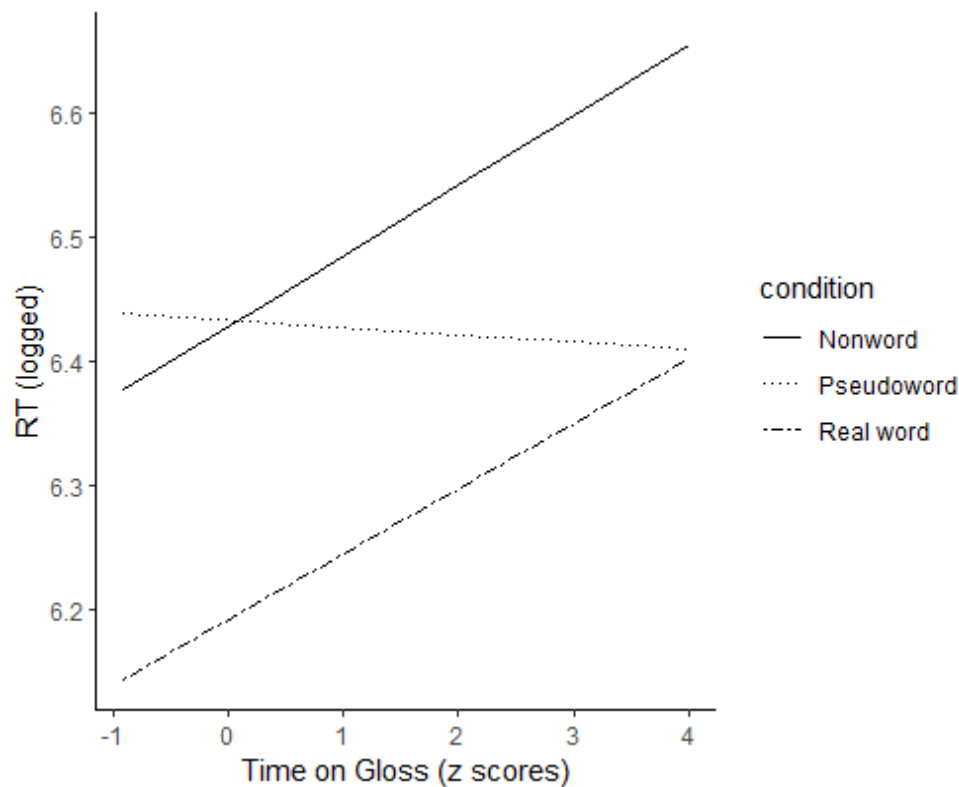
Table 23 presents results for position 2. At this position, the L2 gloss group showed significantly shorter RTs when responding to the real word condition than to the pseudoword condition. RTs were similar in the nonword and the pseudoword conditions. Time on gloss significantly moderated RT difference among the three conditions (see Figure 20): greater time on gloss saw slower responses in the nonword and the real word conditions, while RTs in the pseudoword condition remained relatively constant (i.e., no significant main effect of time on gloss).

Table 23*Self-paced reading delayed posttest: Position 2, L2 Gloss Group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.43 [6.36, 6.50]	.03	189.26	<.001***	.03	.16	.003	.06
Nonword	-.007 [-.06, .05]	.03	-.24	.81				
Real word	-.24 [-.30, -.19]	.03	-8.64	<.001***				
FoO	.009 [-.02, .04]	.02	.54	.60				
Vocabulary size	-.01 [-.06, .04]	.03	-.41	.69				
Time on gloss	-.0002 [-.05, .05]	.03	-.01	.99				
Nonword * Time on gloss	.06 [.01, .12]	.03	2.16	.03*				
Real word * Time on gloss	.06 [.003, .12]	.03	2.09	.04*				

Figure 20

*Condition * Time on Gloss Interaction (Position 2, Delayed Self-paced Reading Test, L2 Gloss Group)*



Summary of self-paced reading results. Regardless of position, test timing, and group, RTs in the pseudoword condition were in general significantly slower than in the real word condition, indicating processing difficulties during reading of the pseudowords. RTs in the pseudoword condition were similar to those in the nonword condition, except at position 2 in the immediate posttest for the L1 gloss group, where RTs in the pseudoword condition were significantly slower than in the nonword condition. This RT pattern suggested that pseudowords were read in a similar manner to nonwords, when other variables were at their mean standardized values.

The number of times a target word appeared in the reading (i.e., FoO) seemed to have facilitated the L1 gloss group's subsequent processing of the pseudowords in the self-paced

reading tests: in the immediate posttest at position 2 (see Table 17 and Figure 15 for the Condition * FoO interaction); and in the delayed posttest at the critical position (see Table 20 and Figure 18). No other variables (i.e., vocabulary size and time on gloss) showed a facilitative effect on pseudoword processing for the L1 gloss group. The L2 gloss group's processing of pseudowords was not affected by any variables included in the analyses.

Thus, to answer RQ2s a and b, when target words appeared a certain number of times, and when participants spent sufficient time on glosses, the L1 gloss group was able to process the pseudowords at a speed similar to real words, indicating fluent retrieval of the pseudowords. For the L2 gloss group, pseudowords were always processed similarly to nonwords and significantly slower than to real words, regardless of target word FoO, time on reading the glosses, and vocabulary size. The difference between the L1 and L2 gloss groups suggested that there might be a gloss language effect, in that with certain conditions, pseudowords learned through L1 but not L2 glosses may be retrieved as fluently as real words.

Summary of Key Significant Effects

Table 24 presents the key significant effects in gloss engagement, meaning recall, meaning matching, and self-paced reading tests.

Table 24
Key Significant Effects

Dependent variables	Significant effects	Interpretation
Gloss processing (binary)	Group (L1 gloss vs. L2 gloss)	L1 gloss group was more likely to skip a gloss than the L2 gloss group.
	Vocabulary size	For both groups, learners with a larger vocabulary size were less likely to skip a gloss.
Immediate meaning recall accuracy	Group * Time on gloss	Time on gloss had a larger, positive effect on L1 gloss group than on the L2 gloss group (see Figure 12).
Delayed meaning recall accuracy	Group * FoO * Time on gloss	L2 gloss group benefitted more from greater time on gloss than the L1 gloss group when FoO was high (see Figure 13).
Immediate meaning matching accuracy	FoO	Both groups were more likely to respond to the meaning matching items correctly as FoO increased.
	Time on gloss	Both groups were more likely to respond to the meaning matching items correctly as time spent on glosses increased.

Table 24 (cont'd)

	Vocabulary size	In both groups, learners with larger vocabulary size were more likely to respond to the test items correctly.
Delayed meaning matching accuracy	Group * Time on gloss	Time on gloss had a larger, positive effect on L1 gloss group than on the L2 gloss group (see Figure 14).
Immediate self-paced reading, L1 gloss group, RTs	Condition (pseudoword, nonword, real word) * FoO	At position 2, RTs in the pseudoword condition were shorter and closer to RTs in the real word condition, as FoO increased (see Figure 15).
Delayed self-paced reading, L1 gloss group, RTs	Condition * Time on gloss * FoO	At the critical position, RTs for the pseudoword decreased with increasing time on gloss, only when FoO was high (see Figure 18).

CHAPTER 4: DISCUSSION AND CONCLUSION

The goal of the current study was to look closely at the gloss language effect, and the factors that moderated the gloss language effect on learners' engagement with glosses and on vocabulary learning from reading. Specifically, the study attempted to answer the following questions: How do learners reading first language (L1) and second language (L2) glosses engage with the glosses (RQ1a)? How does learners' vocabulary size affect their differential engagement with L1 and L2 glosses (RQ1b)? How does gloss language affect learning, both immediately and with a two-week interval after reading, in terms of receptive meaning matching, productive meaning recall, and fluency of online lexical retrieval (RQ2a)? To what extent is the gloss language effect on learning moderated by target words' frequency of occurrence (FoO), learners' vocabulary size, and learners' engagement with the glosses (RQ2b)?

In what follows, I first discuss gloss language effect on learners' engagement with the glosses (RQs1a & b). I then talk about findings regarding gloss language effect on learning (RQs2a & b). In particular, I focus on gloss language effect on the learning of different aspects of word knowledge, and whether and how the variables of FoO, time on gloss, and vocabulary size moderated the gloss language effect. Finally, based on the findings, I discuss the implications for L2 vocabulary pedagogy, bilingual lexicon theories, and gloss language research methodology. I also reflect on the limitations of the current study and suggest directions for future research.

Gloss Language Effect on Gloss Engagement

Gloss engagement was affected by gloss language in that the L1 gloss group was significantly more likely to ignore glosses than the L2 gloss group. To take a closer look at why participants ignored a gloss, I did a follow-up analysis on participants' responses in the exit

questionnaire. Specifically, I looked at Question 5, where participants indicated, on a 100-point scale, the percentage of time they skipped a gloss due to each of the four reasons given in the questionnaire. Table 25 summarizes the results for this question. For both groups, most of the time, they skipped a gloss because they “have guessed the word meaning”, followed by the reasons “I didn’t need the gloss to understand the reading”, and “I didn’t think knowing word meanings was important”. Participant seldom ignored a gloss because “the glosses were not helpful”. Mann Whitney U tests revealed that there was significant group difference in the responses to the first reason (“I have guessed the word meaning”) ($W = 1866.50, p = .02, r = .23$). All other reasons did not see significant differences between the two gloss language groups.

Table 25
Summary of reasons for skipping glosses

Reasons for gloss skipping	Mean percentage of time (<i>SD</i>)	
	L1 gloss group	L2 gloss group
“I have guessed the word meaning.”	33.62 (35.66)	18.04 (27.24)
“I didn’t need the gloss to understand the reading.”	26.52 (34.84)	15.58 (26.83)
“I didn’t think knowing word meanings was important.”	13.98 (26.83)	7.92 (18.37)
“The glosses were not helpful.”	3.63 (13.06)	4.42 (13.11)

It seems that the L1 gloss group was more successful or perceived themselves to be more capable of guessing the meanings of unfamiliar words. This may partially explain why the L1

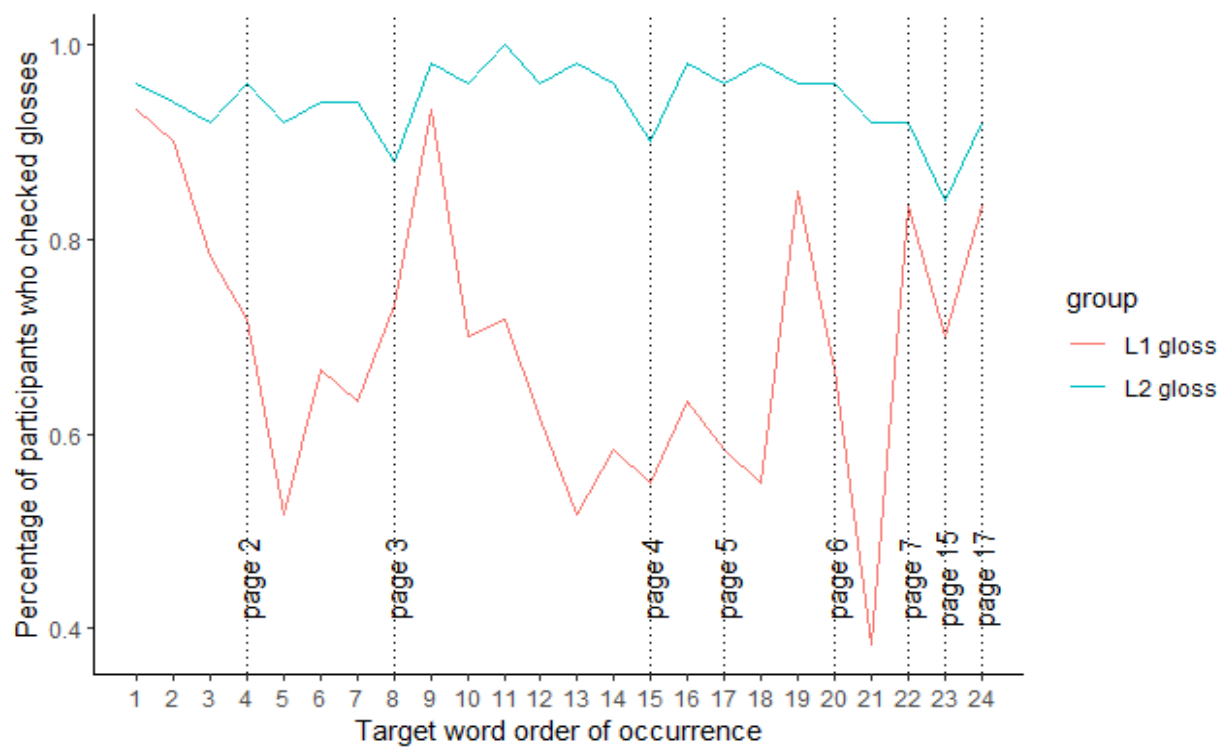
gloss group tended to skip glosses more so than the L2 gloss group. One hypothesis for why the L1 gloss group was more confident in their word guessing ability is that glosses written in the L1 might have given those participants a sense that the target words were easy to understand because it was easier to map the L1 glosses to familiar L1 words. Glosses written in the L2, on the other hand, might be harder to map onto familiar words or concepts in the L1. Participants in the L2 gloss group may thus have thought that the glosses were describing new concepts and that it was difficult for them to get the meanings of the target words by just guessing.

Another reason why the L1 gloss group skipped more glosses might be due to the greater interruption L1 glosses imposed. The findings in Varol and Erçetin (2021) were somewhat similar to those in the current study regarding gloss engagement: significant difference was found in gloss access frequency but not gloss time based on the position of a gloss. The authors in the study attributed the difference in gloss access frequency to one type of gloss being more interrupting in the reading than the other. In the current study, L1 glosses may interrupt the reading flow more because participants had to switch back and forth between their L1 and L2. If L1 glosses caused more interference in reading, the L1 gloss group might have been less likely to check a gloss for target words that appeared later in the reading. Figure 21 plots the percentage of participants who checked a gloss by the order of appearance of target words. The vertical dotted lines represent the beginning of a new page. Right before page 6 (the 20th target word) and page 15 (the 23rd target word), learners had a break in reading, during which they answered two comprehension questions. One striking feature in the figure is that the percentage of gloss-checking learners in the L2 gloss group remained relatively stable throughout reading, while the gloss checking pattern of the L1 gloss group had more ups and downs. Overall, it seems that the L1 gloss group showed a tendency of less gloss checking as time went by: the three spikes at the

19th, 22nd, and 24th target words were all lower than the spikes at the 1st and the 9th target words. In addition, the intervals between the spikes became slightly larger, i.e., longer periods of low percentages in gloss checking, before the L1 gloss group encountered the first set of comprehension questions (the 20th target word). The first interval was eight words between the first and the ninth target words, and the second interval was 10 words between the ninth and the 19th target words. After the first set of comprehension questions, the spikes were more frequent, i.e., shorter periods of low percentages in gloss checking. It was like the L1 gloss group had a ‘reset’, starting to actively access glosses again after the comprehension questions. However, the exact reasons why the L1 gloss group behaved this way is not known. Interview and think-aloud data are needed to understand more accurately the gloss access pattern of the L1 gloss group.

Figure 21

Gloss Processing by Target Word Occurrence Order



Among those who spent enough time processing the glosses, there was no significant group difference in gloss reading time. This reflects that the L1 and L2 glosses may have attracted similar amount of attention or have posed a similar amount of reading demand. In other words, even when glosses were written in the L2, participants may not have found them harder to understand than those written in the L1. The lack of group difference in gloss time was similar to the finding in Warren et al. (2018), where no significant difference in total reading time from eye-tracking data was found among three types of glosses, i.e., text, picture, and multimodal. Findings in the current study, Varol and Erçetin (2021), and Warren et al. (2018) indicate that attention or engagement differences for different types of glosses may exist but may be very small and insignificant. In reality, regardless of the format or modality of the glosses, L2 learners are likely to pay attention to glosses when available, without being drawn to a particular type of gloss. It is interesting to note that the lack of difference in gloss reading time did not lead to the lack of effect of this variable on learning in the current study. The L1 gloss group seemed to have gotten more out of every second spent with the glosses, resulting in better learning performance in some posttests. I discuss this in more detail in the next section.

Gloss Language Effect on Learning

In this section, I first compare gloss language effect on different aspects of learning, namely the development of receptive and productive word meaning knowledge, and fluency and of lexical retrieval. I then compare the effects of FoO, time on gloss, and vocabulary size on the gloss language effect and discuss the different paths of memory consolidation through the L1 and the L2 glosses.

Aspects of Learning

The literature review reveals contradictory findings among previous gloss language studies, with some reporting advantages of L1 or L2 glosses and some finding no language effect. Many of the previous studies, however, did not control or at least report the comparability of the L1 and L2 glosses in terms of length, and whether the words used in the L2 glosses were likely to be known by learners. In the current study, after carefully matching the length of the L1 and L2 glosses, and using L2 words in the glosses likely to be familiar to participants as indicated by their vocabulary size, it was found that short-term development of receptive meaning knowledge measured by an immediate meaning matching test was the only aspect of learning not affected by gloss language. For this aspect of word knowledge, the two gloss groups' performance was similar in the posttest, which was positively associated with longer gloss reading time, higher FoO, and larger vocabulary size. However, neither gloss group benefitted more from any of the three elements.

Other aspects of learning showed a gloss language effect under certain circumstances, revealing a sophisticated picture of how the number of times participants encountered a target word and the amount of time participants spent on reading glosses jointly shaped the gloss language effect on learning. For the retention of receptive meaning knowledge and short-term development of productive meaning knowledge, the L1 gloss group benefitted more from longer gloss reading time while the L2 gloss group was barely affected by the variable (see Figure 12 & Figure 14). The interactions of gloss time and gloss language found in the delayed meaning matching and immediate meaning recall posttests indicate that when being tested on delayed receptive knowledge retention, which was more demanding than short-term development, and on productive knowledge, which was arguably harder to develop (e.g., Nation, 2019; Pellicer-

Sánchez, 2016), time spent on the L1 was ‘worth more’ than with the L2. It is possible that L1-L2 pairs (i.e., L1 glosses and L2 word forms) were easier to register in memory than L2-L2 pairs (i.e., L2 glosses and L2 word forms), and these traces of memory may be effectively enhanced with time on gloss. Such advantage of the L1 did not show up in the learning of word knowledge that was easier to obtain, such as short-term receptive meaning knowledge.

Interestingly, when test demand was even higher, i.e., when the retention of productive knowledge was assessed, the L2 glosses showed an advantage, under the circumstance of long gloss reading time and high FoO combined. For this aspect of learning, the L1 gloss group only benefitted more from long gloss reading time when FoO was low, and this benefit gradually became smaller as FoO increased. The L2 gloss group eventually overtook the L1 gloss group with long gloss reading time and high FoO. This three-way interaction of gloss language, FoO, and gloss time suggests that learning words through the L2 may have an advantage for the retention of productive word knowledge only when (1) the initial processing of word meanings was deep (as shown by long gloss time), and (2) the word meanings were subsequently reinforced enough times through encountering the words again. The interaction also indicated that it may take more time for the advantage of L2 glosses to show up, i.e., in the delayed but not immediate posttest. In addition, it seems that when it comes to long-term retention of productive meaning knowledge, the memory of L1-L2 pairs was not strengthened by subsequent encounters with the words and was even slightly reduced by those encounters.

Fluency of lexical retrieval is the hardest aspect of word knowledge to obtain and often takes longer to develop than declarative word knowledge measured in offline tests. Several previous studies have shown that even when learners have mastered receptive and/or productive declarative word knowledge, they may not be able to establish retrieval fluency for the newly

learned words (e.g., Y. Chen, 2021; Elgort & Warren, 2014). Previous gloss language studies rarely examined the retrieval fluency of newly learned target words. In the current study, albeit not straightforwardly, a clear gloss language effect can be observed in fluency of retrieval indicated by the RT patterns in the self-paced reading immediate and delayed posttests. Under no circumstances in the current study did the L2 gloss group react to the target words in a similar manner to the real words. The L1 gloss group showed some degree of retrieval fluency in the immediate posttest for target words with high FoOs and in the delayed posttest for words with a high FoO and long gloss time combined. Again, findings in the self-paced reading tests show that when test demand was higher, such as in the delayed posttests, a combination of initial attention to gloss and subsequent reinforcement through repeated encounters is needed for learning to take place.

Findings of lexical retrieval fluency were relevant to the development of the bilingual lexicon. The Literature Review introduces three bilingual lexicon models, all of which predict that L2 vocabulary learning through the L1 may reduce the chance of establishing a direct conceptual link for an L2 word and thus result in worse quality of word knowledge. Although the current study did not directly test whether the target words were directly linked to their concepts, fluency of retrieval was examined, which could, to some extent, indirectly reflect the strength of the links between the target words and their concepts, and thus the quality of word knowledge. Results show that after reading a fictional story in two days, only participants who learned the target words through L1 input were able to develop some degree of retrieval fluency, which does not seem to align with the prediction of the bilingual lexicon models and is in contrast with findings in intentional vocabulary learning studies in this paradigm. This may have to do with the differences in the nature of learning in intentional and incidental contexts. Learners encounter L2

words in relatively more varied and richer contexts in incidental learning. The rich context where the L2 words are embedded may make the differences in learning through the L1 and L2 smaller in this aspect of word knowledge. In addition, appearances of a word may be more spread out in incidental learning, which may reduce the benefits of learning through the L2 in creating quality lexical representations. Unlike in intentional learning, where lexical representations established through L2 input can be reinforced multiple times within a very short interval, in incidental learning, reinforcement comes sporadically and may not be in time before the memory created through L2 input fades.

However, caution should be taken here not to overly interpret the findings in relation to bilingual lexicon development. First, fluency of retrieval is not equivalent to direct conceptual links. Even when target words learned through the L1 were retrieved somewhat fluently, it did not mean that the target words were directly linked to the concepts. It could be the case that only the L1-L2 form-form connections were strengthened during learning. This is akin to stage 2 in the model proposed by Jiang (2000), where fast L1-L2 retrieval is achieved. Second, learning time and number of exposures in the current study may not have allowed the L2 gloss group to develop lexical retrieval fluency. Given the time and opportunity of repeated exposure, the L2 gloss group may have performed differently.

To summarize, findings in the current study reveal intricate gloss language effects on learning, varied across the aspects of knowledge measured. First, L1 and L2 glosses benefitted the learning of different aspects of word knowledge under different circumstances: L1 gloss could contribute more to the retention of receptive meaning knowledge, short-term development of productive meaning knowledge, and fluency of lexical retrieval; L2 gloss was found to facilitate the long-term retainment of productive meaning knowledge. These findings also

indicate that the benefits of L2 gloss were more likely to be found when the type of knowledge tested was harder to develop. Second, it seems that it would take longer for the advantage of L2 gloss to show up, i.e., in the delayed posttests. In the same vein, one explanation for why the L2 gloss group failed to retrieve target words as fluently as real words in the self-paced reading posttests may be that the participants needed longer time to integrate glosses written in the L2.

Effects of Moderating Variables

Most previous gloss language studies simply compared the L1 and L2 gloss groups, without considering potential moderating factors. The current study included three variables, namely target word FoO, time spent on reading glosses, and participants' vocabulary size. These variables had differential effects on how L1 and L2 glosses affected learning, moderating the gloss language effect.

Vocabulary size was the only variable that had a negligible moderating effect on the comparison of L1 and L2 glosses. Nor did vocabulary size influence learning gains much. The main effect of vocabulary size was found only in the meaning matching posttests, both immediate and delayed. The lack of effect of vocabulary size on the learning of most aspects of word knowledge aligns with the result in Yanagisawa et al.'s (2020) meta-analysis and was probably because glosses reduced the need to guess word meanings from context, for which vocabulary size may be an important factor. The lack of moderating effect of vocabulary size on gloss language may be attributed to the familiarity of words used in the L2 glosses. All words used in the L2 glosses were within the 3k frequency range, which the participants were likely to know based on their vocabulary size test scores. It seems that once a lexical threshold was reached, or that comprehension of glosses was not an issue, vocabulary size may not have a large effect on gloss language. An alternative perspective to look at the null moderating effect of

vocabulary size is that knowing more words in general would not facilitate participants in reading either gloss types or in integrating word meanings the glosses provided. In contrast, spending more time on glosses and/or encountering the words repeatedly were more likely to give participants a better chance to learn the target words, which I discuss in the next paragraphs.

According to the Levels of Processing framework (Craik & Lockhart, 1972), time on gloss, or the amount of engagement with the glosses, reflects the depth of processing, which predicts the strength of memory. There were three key findings regarding the moderating effects of time on gloss. First, as mentioned in the last section, the L1 gloss group benefitted more from longer gloss reading time than the L2 gloss group (i.e., a group and time on gloss interaction). In other words, the same level of processing depth through the L1 and the L2 glosses did not result in the same level of memory strength. Such finding suggests that information processed in the L1 may leave a stronger memory trace. The second key finding further corroborates this argument: the amount of time on gloss hardly moderated the effects of L2 glosses. It seems that deeper levels of processing in the L2 did not guarantee improved memory strength of target words. Relatedly, the third key finding reveals that depth of processing as reflected in the amount of gloss time facilitated learning through L2 glosses only when FoO was high. In addition, the positive effect of gloss time on the development of lexical retrieval fluency was also only present for target words with high FoOs. In these situations, depth of processing was not enough but had to be paired with subsequent reinforcement through repeated exposure in order to improve learning. The combined effects of gloss time and FoO are discussed in more detail later.

The effect of FoO, or repeated exposure to target words, also had to do with memory strength. As mentioned in the Literature Review, the instance-based models for word learning (Bolger et al., 2008; Reichle & Perfetti, 2003) hypothesized that the incomplete word knowledge

resulting from the initial encounter with a word is reactivated in subsequent encounters, until full word knowledge is extracted. FoO can thus be seen to represent memory reactivation and reinforcement. Unlike depth of processing that is represented by the amount of time on glosses, FoO, or memory reinforcement, had positive effects on both the L1 and L2 gloss groups, i.e., lack of group and FoO interactions, in the learning of productive and receptive meaning knowledge. It means that, in these aspects of learning, regardless of the initial depth of processing of the glosses, memory traces of target words learned through the L1 and L2 glosses would be reinforced to a similar degree. In the short-term development of lexical retrieval fluency, however, FoO only had an effect on the L1 gloss group. Such finding indicates that within the current range of FoO, that is, within the current level of memory reinforcement provided, only memory traces created through the L1 may be reinforced to the extent where lexical retrieval fluency can be developed. Memory traces created through the L2 may need more instances of reactivation and reinforcement to reach the same level of strength as those created through the L1. A possible reason is that memory traces in the L1 might be easier to retrieve and thus reactivate. FoO was also the only variable that had an impact on short-term lexical retrieval fluency development. Reading time on glosses on the other hand, did not affect this aspect of learning. This shows that for the retrieval fluency aspect of word knowledge, repeated reactivation of a word may be more important than its initial depth of processing.

The three-way interaction of group, FoO, and time on gloss was found in the meaning recall delayed posttest for the L2 gloss group and the self-paced reading delayed posttest for the L1 gloss group. One thing to note in this discussion is that glosses were only provided in the first appearance of a target word. The indication of this design is that the depth of processing achieved through spending time on reading the glosses only refers to the depth of initial

processing. In other words, gloss time was most pertinent to the initial registration, but not or at least not directly relevant, to depth of subsequent processing. These three-way interactions show that initial depth of processing or subsequent memory reactivation alone was insufficient for some aspects of learning, especially those hard to develop, i.e., in this case long-term retention of productive knowledge and retrieval fluency. For the L2 gloss group, the depth of processing of L2 glosses facilitated learning only when the words were subsequently reactivated multiple times. The need for combined effects of FoO and gloss time may be due to greater difficulty in memorizing the L2 glosses. Therefore, the initial registration of the L2 glosses needed to be reactivated in context again and again for learning to take place. In addition, the L2 gloss group's learning of this aspect of word knowledge became larger than the L1 gloss group's. One hypothesis is that because L2 glosses were harder to memorize, the L2 gloss group deliberately tried to retrieve the L2 gloss for a target word in each subsequent encounter, resulting in stronger memory and thus better performance in tests for declarative word knowledge. In comparison, L1 glosses were easier to memorize, giving the L1 gloss group a sense of confidence (see also the section on Gloss Engagement) that they remembered the glosses well. As a result, the L1 gloss group may not have consciously (but may be subconsciously so) tried to retrieve the glosses when seeing the target words.

The challenge of maintaining lexical retrieval fluency also required both deep initial processing and subsequent memory reinforcement: only when there were enough opportunities to reinforce the initial memory traces, would deeper level of initial processing of L1 glosses be facilitative for fluency retention. Again, the L2 gloss group did not achieve the long-term lexical retrieval fluency as the L1 gloss group, possibly because the L2 gloss group may take longer time or need more target word exposures to do so. It is interesting to note that in the meaning

recall delayed posttest, the three-way interaction also reveals that with repeated encounters with the target words, the effect of initial processing depth on the L1 gloss group was reduced. This contrasts with results in the self-paced delayed posttest, where FoO and time on gloss together facilitated word retrieval of the L1 group. One tentative explanation could be that the ease of processing of L1 glosses, the great depth of processing (i.e., long gloss reading time), and repeated activation (i.e., high FoOs), created rich and detailed semantic representations, which became a burden for conscious memory retrieval while facilitating online processing of the target words, resulting in the differential combined effect of FoO and gloss time on productive knowledge and retrieval fluency.

Implications

The current study has pedagogical, theoretical, and methodological implications. Pedagogically, findings in the current study show that L1 and L2 glosses benefited different aspects of vocabulary learning. Language instructors, material writers, and ed tech designers should choose gloss language adaptively, based on the aspects of vocabulary knowledge learners most need for certain words, and the amount of exposure learners are likely to have to the words. For example, if learners only need short-term receptive meaning knowledge for some words, L1 and L2 glosses are likely to have similar effects, giving language practitioners more freedom to choose gloss language, depending on the target reader population's L1 backgrounds (i.e., miscellaneous or homogenous). Or, if the goal is to develop lexical retrieval fluency and the learning phase is short, L1 glosses are preferred. Second, whichever aspect of word knowledge is targeted, the key to more successful learning is to encourage learners to engage with the glosses and to increase the opportunities learners encounter the words. Longer engagement with glosses may be particularly useful when the glosses are written in the L1. When using L2 glosses is the

preferred choice, e.g., for learners from different L1 backgrounds, repeated exposure is the key to allow learners reap the benefits of glosses, based on the result that only when FoO was high, did the L2 gloss group benefit from greater gloss engagement.

Results in the current study support the bilingual teaching approach, i.e., the use of L1 and L2 together, instead of using the L2 exclusively. Within the time limit of the learning phase, i.e., two days, participants who read the L1 glosses did better in more aspects of learning than those who read the L2 glosses, given the right conditions. It could be the case, though, that the advantages of L2 glosses take longer to show up. In any case, the key message here is that the use of L1 should not be seen as a barrier and can be used to support L2 learning. Word learning through the L1 may be more efficient. This was the case even when learners have reached a certain proficiency threshold to allow them to comprehend the L2 glosses without issues.

Theoretically, the findings of the current study shed some light on the development of bilingual lexicon. Within a short learning phase (i.e., two days), L1 glosses, but not L2 glosses, were more efficient in facilitating the development of lexical retrieval fluency. In other words, the L2 glosses did not lead to the establishment of direct conceptual links for the newly learned words. However, it is hard to pinpoint whether retrieval fluency shown by the L1 gloss group came from the establishment of direct conceptual links of the newly learned words, or merely from the improved automaticity of L1-mediated access to concepts.

In terms of methodological implication, the study demonstrates the use of hyperlinks to track learners' gloss engagement. Although hyperlinks as a tracking tool have been used in other research areas (see H. Lee & Lee, 2013), they have seldom been used in gloss language research. With online data collection getting more common, hyperlinks offer a convenient alternative to eye tracking in monitoring engagement and attention during learning.

Limitations and Future Directions

While the study is one of the first to examine potential moderating factors on the gloss language effect on both learning and gloss engagement, it is not without limitations. First, gloss time tracked through hyperlinks may not have accurately reflected every participant's engagement with glosses. While participants were instructed to click close a gloss when they finished reading it, some participants may have delayed doing so, resulting in the gloss reading time recorded being longer than the actual reading time. One solution may be to implement timed glosses, which appear for a specific amount of time, depending on gloss length. After a gloss disappears, participants have to re-click the gloss to read it again. Although this way may reduce the likelihood of participants forgetting to close the gloss, attention tracking using hyperlinks cannot be compared with eye tracking. But hyperlinks have the advantage of allowing data collection online. This is a tradeoff researchers need to consider.

Related to online data collection, one concern regards how to make sure that learners are paying attention during the experiment. The reading comprehension questions during learning and the comprehension questions in the self-paced reading tests have served as attention checks, and participants who did not reach a certain comprehension level were excluded. Still, future studies could insert more attention checks, such as by asking participants to click on a particular photo out of many displayed on screen. These non-comprehension-based attention checks can add a further layer of security against distraction.

Finally, the distribution of FoO was imbalanced, with more target words in the FoO range between 1 and 10, and fewer above 10. The smaller number of words with an FoO above 10 means that data points were sparse for this FoO range, leading to less accurate estimates of the data. The uneven distribution of FoO was due to restrictions in the original graded reader. Future

studies may have to implement more adaptations to increase the number of target words for certain FoO ranges.

Besides fixing the above issues, future research would also benefit from examining the long-term gloss language effect by implementing delayed posttests with a longer interval after the learning phase. The range of FoO should also be increased. Such designs allow us to gain insights into how the gloss language effect changes overtime, e.g., L2 gloss may benefit retrieval fluency more as time goes by and as learners are exposed to words more times. In addition, the current study shows differences in gloss checking but not gloss reading time. Interviews can be conducted to further investigate why this may be the case. Finally, replication studies should be conducted with learners from other L1 backgrounds, proficiency levels, and age. Learners' familiarity with technology should also be considered, which may have an impact on how learners engage with glosses on a digital device.

Conclusions

The study reveals a sophisticated picture of how the L1 and L2 glosses may benefit the learning of different aspects of word knowledge under various circumstances as influenced by the number of repeated exposures, gloss engagement, and learners' vocabulary size. Findings in the study bear pedagogical implications for language instructors and material writers as to when to use which type of gloss, based on the learning context and learners' needs. More importantly, the study shows that L1 is an asset L2 learners can take advantage of, instead of something to be avoided in L2 learning. The study also contributes to theories regarding the bilingual model lexicons and the use of hyperlinks as a tracking tool alternative to eye tracking.

REFERENCES

- American Council on the Teaching of Foreign Language (ACTFL). (n.d.). *Facilitate Target Language Use*. <https://www.actfl.org/educator-resources/guiding-principles-for-language-learning/facilitate-target-language-use>
- Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, 21(3), 199–226. <https://doi.org/10.1080/09588220802090246>
- Altarriba, J., & Mathis, K. M. (1997). Conceptual and lexical development in second language acquisition. *Journal of Memory and Language*, 36(4), 550–568. <https://doi.org/10.1006/jmla.1997.2493>
- Antón, M., & DiCamilla, F. (1998). Socio-cognitive functions of L1 collaborative interaction in the L2 classroom. *The Canadian Modern Language Review*, 54(3), 314–342. <https://doi.org/10.3138/cmlr.54.3.314>
- 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. <https://doi.org/10.1016/j.jml.2012.11.001>
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459. <https://doi.org/10.3758/BF03193014>
- Bartolotti, J., & Marian, V. (2017). Bilinguals' existing languages benefit vocabulary learning in a third language. *Language Learning*, 67(1), 110–140. <https://doi.org/10.1111/lang.12200>
- Boers, F. (2022). Glossing and vocabulary learning. *Language Teaching*, 55(1), 1–23. <https://doi.org/10.1017/S0261444821000252>
- Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, 66, 113–129. <https://doi.org/10.1016/j.system.2017.03.017>
- Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45(2), 122–159. <https://doi.org/10.1080/01638530701792826>
- Brevik, L. M., & Rindal, U. (2020). Language use in the classroom: Balancing target language exposure with the need for other languages. *TESOL Quarterly*, 54(4), 925–953. <https://doi.org/10.1002/tesq.564>

- Brooks-Lewis, K. A. (2009). Adult learners' perceptions of the incorporation of their L1 in foreign language teaching and learning. *Applied Linguistics*, 30(2), 216–235. <https://doi.org/10.1093/applin/amn051>
- Brown, A. (2021). Monolingual versus multilingual foreign language teaching: French and Arabic at beginning levels. *Language Teaching Research*, 1362168821990347. <https://doi.org/10.1177/1362168821990347>
- Brown, A., & Lally, R. (2019). Immersive versus nonimmersive approaches to TESOL: A classroom-based intervention study. *TESOL Quarterly*, 53(3), 603–629. <https://doi.org/10.1002/tesq.499>
- Bruen, J., & Kelly, N. (2017). Using a shared L1 to reduce cognitive overload and anxiety levels in the L2 classroom. *The Language Learning Journal*, 45(3), 368–381. <https://doi.org/10.1080/09571736.2014.908405>
- Bruton, A., López, M. G., & Mesa, R. E. (2011). Incidental L2 vocabulary learning: An impracticable term? *TESOL Quarterly*, 45(4), 759–768. <https://doi.org/10.5054/tq.2011.268061>
- Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109, 104047. <https://doi.org/10.1016/j.jml.2019.104047>
- Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition*, 13(3), 359–371. <https://doi.org/10.1017/S1366728909990344>
- Canagarajah, S. (2011). Translanguaging in the classroom: Emerging issues for research and pedagogy. *Applied Linguistics Review*, 2, 1–28. <https://doi.org/10.1515/9783110239331.1>
- Canagarajah, S. (2013). *Translingual Practice: Global Englishes and Cosmopolitan Relations*. Routledge.
- Carless, D. (2007). Student use of the mother tongue in the task-based classroom. *ELT Journal*, 62(4), 331–338. <https://doi.org/10.1093/elt/ccm090>
- Celce-Murcia, M. (2014). An overview of language teaching methods and approaches. *Teaching English as a second or foreign language*, 4, 2–14.
- Chambers, F. (1991). Promoting use of the target language in the classroom. *Language Learning Journal*, 4(1), 27–31.
- Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge University Press.

- Chen, B., Ma, T., Liang, L., & Liu, H. (2017). Rapid L2 word learning through high constraint sentence context: An event-related potential study. *Frontiers in Psychology*, 8, 2285. <https://doi.org/10.3389/fpsyg.2017.02285>
- Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, 31(5), 693–713. <https://doi.org/10.1093/applin/amq031>
- Chen, Y. (2021). Comparing incidental vocabulary learning from reading-only and reading-while-listening. *System*, 97, 102442.
- Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences*, 45, 137–143. <https://doi.org/10.1016/j.lindif.2015.11.018>
- Choi, S. (2017). Processing and learning of enhanced English collocations: An eye movement study. *Language Teaching Research*, 21(3), 403–426. <https://doi.org/10.1177/1362168816653271>
- Chun, D. M., & Payne, J. S. (2004). What makes students click: Working memory and look-up behavior. *System*, 32(4), 481–503. <https://doi.org/10.1016/j.system.2004.09.008>
- Cobb, T. (n.d.). *Vocabprofile* [Computer program]. Retrieved April 30, 2021, from <https://www.lex tutor.ca/vp/>.
- Comesaña, M., Perea, M., Piñeiro, A., & Fraga, I. (2009). Vocabulary teaching strategies and conceptual representations of words in L2 in children: Evidence with novice learners. *Journal of Experimental Child Psychology*, 104(1), 22–33. <https://doi.org/10.1016/j.jecp.2008.10.004>
- Cook, G. (2010). *Translation in language teaching: An argument for reassessment*. Oxford University Press.
- Cook, V. (2001). Using the first language in the classroom. *The Canadian Modern Language Review*, 57(3), 402–423. <https://doi.org/10.3138/cmlr.57.3.402>
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The Modern Language Journal*, 94(1), 103–115. <https://doi.org/10.1111/j.1540-4781.2009.00986.x>

- Creese, A., & Blackledge, A. (2015). Translanguaging and identity in educational settings. *Annual Review of Applied Linguistics*, 35, 20–35. <https://doi.org/10.1017/S0267190514000233>
- Cummins, J. (2007). Rethinking monolingual instructional strategies in multilingual classrooms. *Canadian Journal of Applied Linguistics*, 10(2), 221–240.
- De Costa, P. I., Singh, J. G., Milu, E., Wang, X., Fraiberg, S., & Canagarajah, S. (2017). Pedagogizing translanguaging practice: prospects and possibilities. *Research in the Teaching of English*, 51(4), 464–472.
- De la Campa, J. C., & Nassaji, H. (2009). The amount, purpose, and reasons for using L1 in L2 classrooms. *Foreign Language Annals*, 42(4), 742–759. <https://doi.org/10.1111/j.1944-9720.2009.01052.x>
- DiCamilla, F. J., & Antón, M. (2012). Functions of L1 in the collaborative interaction of beginning and advanced second language learners. *International Journal of Applied Linguistics*, 22(2), 160–188. <https://doi.org/10.1111/j.1473-4192.2011.00302.x>
- Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3), 175–197. <https://doi.org/10.1017/S1366728902003012>
- Dijkstra, T., van Heuven, W. J. B., & Grainger, J. (1998). Simulating cross-language competition with the bilingual interactive activation model. *Psychologica Belgica*, 38(3–4), 177–196.
- Duff, P. A., & Polio, C. G. (1990). How much foreign language is there in the foreign language classroom? *The Modern Language Journal*, 74(2), 154–166. <https://doi.org/10.2307/328119>
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367–413. <https://doi.org/10.1111/j.1467-9922.2010.00613.x>
- Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition*, 42(1), 7–32. <https://doi.org/10.1017/S0272263119000561>
- Elgort, I., Candry, S., Boutorwick, T. J., Eyckmans, J., & Brysbaert, M. (2018). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*, 39(5), 646–667. <https://doi-org.proxy2.cl.msu.edu/10.1093/applin/amw029>
- Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J. Z. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience*, 30(5), 506–528. <https://doi.org/10.1080/23273798.2014.942673>

- Elgort, I., & Piasecki, A. E. (2014). The effect of a bilingual learning mode on the establishment of lexical semantic representations in the L2. *Bilingualism: Language and Cognition*, 17(3), 572–588. <https://doi.org/10.1017/S1366728913000588>
- Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, 64(2), 365–414. <https://doi.org/10.1111/lang.12052>
- Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, 28(2), 339–368. <https://doi.org/10.1017/S0272263106060141>
- Ender, A. (2016). Implicit and explicit cognitive processes in incidental vocabulary acquisition. *Applied Linguistics*, 37(4), 536–560.
- Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, 24(3), 369–383. <https://doi.org/10.1017/S0142716403000195>
- Fischer, R. (2007). How do we know what students are actually doing? Monitoring students' behavior in CALL. *Computer Assisted Language Learning*, 20(5), 409–442. <https://doi.org/10.1080/09588220701746013>
- Fu, M., & Li, S. (2022). The effects of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition*, 44(1), 2–34. <https://doi.org/10.1017/S0272263120000388>
- Gánem-Gutiérrez, G. A., & Roehr, K. (2011). Use of L1, metalanguage, and discourse markers: L2 learners' regulation during individual task performance. *International Journal of Applied Linguistics*, 21(3), 297–318. <https://doi.org/10.1111/j.1473-4192.2010.00274.x>
- Grange, J. (2015). trimr: An implementation of common response time trimming methods. R package version 1.0.1. <https://CRAN.R-project.org/package=trimr>
- Godfroid, A. (2019). Sensitive measures of vocabulary knowledge and processing: Expanding Nation's framework. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). Routledge.
- Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, 21(3), 563–584. <https://doi.org/10.1017/S1366728917000219>
- 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. <https://doi.org/10.1017/S0272263113000119>

- Gries, S. T. (2021). (Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example. *Language Learning*, 71(3), 757–798. <https://doi-org.proxy2.cl.msu.edu/10.1111/lang.12448>
- Hair Jr, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). Macmillan.
- Hall, G., & Cook, G. (2012). Own-language use in language teaching and learning. *Language Teaching*, 45(3), 271–308. <https://doi.org/10.1017/S0261444812000067>
- Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (1st ed., pp. 258–286). Cambridge University Press. <https://doi.org/10.1017/CBO9781139524780.011>
- Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470756492.ch12>
- 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(3), 327–339. <https://doi.org/10.1111/j.1540-4781.1996.tb01614.x>
- Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the Involvement Load Hypothesis in vocabulary acquisition. *Language Learning*, 51(3), 539–558. <https://doi.org/10.1111/0023-8333.00164>
- Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, 17(1), 19–28. <https://doi.org/10.1111/j.1467-9817.1994.tb00049.x>
- Jeong, H., Sugiura, M., Sassa, Y., Wakusawa, K., Horie, K., Sato, S., & Kawashima, R. (2010). Learning second language vocabulary: Neural dissociation of situation-based learning and text-based learning. *NeuroImage*, 50(2), 802–809. <https://doi.org/10.1016/j.neuroimage.2009.12.038>
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21(1), 47–77. <https://doi.org/10.1093/applin/21.1.47>
- Jiang, N. (2013). *Conducting reaction time research in second language studies*. Routledge.
- Jones, L. C. (2013). Supporting listening comprehension and vocabulary acquisition with multimedia annotations: The students' voice. *CALICO Journal*, 21(1), 41–65. <https://doi.org/10.1558/cj.v21i1.41-65>

- Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228-238.
- Kang, H., Kweon, S.-O., & Choi, S. (2020). Using eye-tracking to examine the role of first and second language glosses. *Language Teaching Research*, 136216882092856. <https://doi.org/10.1177/1362168820928567>
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633. <https://doi.org/10.3758/BRM.42.3.627>
- 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. <https://doi.org/10.1016/j.system.2017.01.009>
- Kim, H. S., Lee, J. H., & Lee, H. (2020). The relative effects of L1 and L2 glosses on L2 learning: A meta-analysis. *Language Teaching Research*, 136216882098139. <https://doi.org/10.1177/1362168820981394>
- Kim, K. M., & Godfroid, A. (2019). Should we listen or read? modality effects in implicit and explicit knowledge. *The Modern Language Journal*, modl.12583. <https://doi.org/10.1111/modl.12583>
- Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly*, 46(1), 56–79. <https://doi.org/10.1002/tesq.3>
- Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, 9(2), 119–135. <https://doi.org/10.1017/S1366728906002483>
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2), 149–174. <https://doi.org/10.1006/jmla.1994.1008>
- Kroll, J. F., van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3), 373–381. <https://doi.org/10.1017/S136672891000009X>
- Kubota, R. (2018). Unpacking research and practice in world Englishes and Second Language Acquisition. *World Englishes*, 37(1), 93–105. <https://doi.org/10.1111/weng.12305>
- Laufer, B. (2005). Instructed second language vocabulary learning: The fault in the 'default hypothesis.' In A. Housen & M. Pierrard (Eds.), *Investigations in instructed second language acquisition*. Mouton de Gruyter. <https://doi-org.proxy2.cl.msu.edu/10.1515/9783110197372>

- Laufer, B. (2006). Comparing focus on form and focus on formS in second-language vocabulary learning. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, 63(1), 149–166. <https://doi.org/10.1353/cml.2006.0047>
- Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, 29(4), 694–716. <https://doi.org/10.1093/applin/amn018>
- Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a call dictionary and how does it affect word retention? *Language Learning & Technology*, 21.
- Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>
- Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it?. *RELC journal*, 28(1), 89–108.
- Lee, H., & Lee, J. H. (2013). Implementing glossing in mobile-assisted language learning environments: Directions and outlook. *Language Learning & Technology*, 17, 6–22.
- Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, 21(2), 32–51.
- Lee, J. H., & Lee, H. (2022). Teachers' verbal lexical explanation for second language vocabulary learning: A meta-analysis. *Language Learning*, 72(2), 576–612.
- Lee, J. H., & Levine, G. S. (2020). The effects of instructor language choice on second language vocabulary learning and listening comprehension. *Language Teaching Research*, 24(2), 250–272. <https://doi.org/10.1177/1362168818770910>
- Lee, J. H., & Lo, Y. Y. (2017). An exploratory study on the relationships between attitudes toward classroom language choice, motivation, and proficiency of EFL learners. *System*, 67, 121–131. <https://doi.org/10.1016/j.system.2017.04.017>
- Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal*, 97(4), 887–901. <https://doi.org/10.1111/j.1540-4781.2013.12044.x>
- Lee, S., & Pulido, D. (2017). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research*, 21(1), 118–135. <https://doi.org/10.1177/1362168816637381>
- Liu, D., Ahn, G.-S., Baek, K.-S., & Han, N.-O. (2004). South Korean high school english teachers' code switching: Questions and challenges in the drive for maximal use of English in teaching. *TESOL Quarterly*, 38(4), 605. <https://doi.org/10.2307/3588282>

- Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, 27(03). <https://doi.org/10.1017/S0272263105050163>
- Loewen, S. (2014). *Introduction to instructed second language acquisition*. Routledge.
- Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. *Foreign Language Research in Cross-Cultural Perspective*, 39.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.
- Lüdecke et al., (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Ma, L. P. F. (2019). Examining the functions of L1 use through teacher and student interactions in an adult migrant English classroom. *International Journal of Bilingual Education and Bilingualism*, 22(4), 386–401. <https://doi.org/10.1080/13670050.2016.1257562>
- Macaro, E. (2001). Analysing student teachers' codeswitching in foreign language classrooms: Theories and decision making. *The Modern Language Journal*, 85(4), 531–548. <https://doi.org/10.1111/0026-7902.00124>
- Macaro, E., & Lee, J. H. (2013). Teacher language background, codeswitching, and english-only instruction: Does age make a difference to learners' attitudes? *TESOL Quarterly*, 47(4), 717–742. <https://doi.org/10.1002/tesq.74>
- Macaro, E., Tian, L., & Chu, L. (2020). First and second language use in English medium instruction contexts. *Language Teaching Research*, 24(3), 382–402. <https://doi.org/10.1177/1362168818783231>
- Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, 40(3), 651–675. <https://doi.org/10.1017/S0272263117000341>
- Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, 39(5), 861–904. <https://doi.org/10.1017/S0142716418000036>
- Miyasako, N. (2002). Does text-glossing have any effects on incidental vocabulary learning through reading for Japanese senior high school students? *Language Education & Technology*, 39, 1–20.

- Mohamed, A. A. (2018). Exposure frequency in L2 reading: An eye-movement perspective of incidental vocabulary learning. *Studies in Second Language Acquisition*, 40(2), 269–293. <https://doi.org/10.1017/S0272263117000092>
- Moore, P. J. (2013). An emergent perspective on the use of the first language in the English-as-a-foreign-language classroom. *The Modern Language Journal*, 97(1), 239–253. <https://doi.org/10.1111/j.1540-4781.2013.01429.x>
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26(2–3), 131–157. <https://doi.org/10.1080/01638539809545042>
- Nakatsukasa, K., & Loewen, S. (2015). A teacher's first language use in form-focused episodes in Spanish as a foreign language classroom. *Language Teaching Research*, 19(2), 133–149. <https://doi.org/10.1177/1362168814541737>
- Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12–25.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Nation, P. (2012). The BNC/COCA word family lists. https://www.wgtn.ac.nz/data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf
- Nation, P. (2019). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp.15–29). Routledge. <http://doi.org/10.4324/9780429291586-28>.
- Nilsson, M. (2020). Beliefs and experiences in the English classroom: Perspectives of Swedish primary school learners. *Studies in Second Language Learning and Teaching*, 10(2), 257–281.
- Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454), 730–745. <https://doi.org/10.1198/016214501753168389>
- Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition *from* and *while* reading: An eye-tracking study. *Studies in Second Language Acquisition*, 38(1), 97–130. <https://doi.org/10.1017/S0272263115000224>
- Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, 22(1), 31–55.
- Peters, E. (2007). Manipulating L2 learners' online dictionary use and its effect on L2 word retention. *Language Learning & Technology*, 11(2), 36–58. <http://dx.doi.org/10125/44103>

- Phillipson, R. (1992). *Linguistic imperialism*. OUP Oxford.
- Polio, C. G., & Duff, P. A. (1994). Teachers' language use in university foreign language classrooms: A qualitative analysis of English and target language alternation. *The Modern Language Journal*, 78(3), 313–326. <https://doi-org.proxy2.cl.msu.edu/10.2307/330110>
- Qian, David. D., & Lin, L. H. F. (2019). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies*. Routledge.
- Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *TESOL Quarterly*, 55(1), 105–133. <https://doi.org/10.1002/tesq.579>
- Rassaei, E. (2020). Effects of mobile-mediated dynamic and nondynamic glosses on L2 vocabulary learning: A sociocultural perspective. *The Modern Language Journal*, 104(1), 284–303. <https://doi.org/10.1111/modl.12629>
- Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading*, 7(3), 219–237. https://doi.org/10.1207/S1532799XSSR0703_2
- Rice, C. A., & Tokowicz, N. (2020). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, 42(2), 439–470. <https://doi.org/10.1017/S0272263119000500>
- Sato, M., & Angulo, I. (2020). The role of L1 use by high-proficiency learners in L2 vocabulary development. In W. Suzuki & N. Storch (Eds.). *Language in language learning and teaching: A collection of empirical studies* (pp.41–66). John Benjamins.
- Sato, M., & Loewen, S. (2018). Metacognitive instruction enhances the effectiveness of corrective feedback: variable effects of feedback types and linguistic targets: Metacognitive instruction and corrective feedback. *Language Learning*, 68(2), 507–545. <https://doi.org/10.1111/lang.12283>
- Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363. <https://doi.org/10.1177/1362168808089921>
- Scott, V. M., & Fuente, M. J. D. L. (2008). What's the problem? L2 Learners' use of the L1 during consciousness-raising, form-focused tasks. *The Modern Language Journal*, 92(1), 100–113. <https://doi.org/10.1111/j.1540-4781.2008.00689.x>
- Shiki, O. (2008). Effects of glosses on incidental vocabulary learning: Which gloss-type works better, L1, L2, single choice, or multiple choices for Japanese university students? *Journal of Inquiry and Research*, 87, 39–56.

- Shin, J.-Y., Dixon, L. Q., & Choi, Y. (2020). An updated review on use of L1 in foreign language classrooms. *Journal of Multilingual and Multicultural Development*, 41(5), 406–419. <https://doi.org/10.1080/01434632.2019.1684928>
- Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: acquisition of collocations under different input conditions. *Language Learning*, 63(1), 121–159. <https://doi.org/10.1111/j.1467-9922.2012.00730.x>
- Storch, N., & Aldosari, A. (2010). Learners' use of first language (Arabic) in pair work in an EFL class. *Language Teaching Research*, 14(4), 355–375. <https://doi-org.proxy2.cl.msu.edu/10.1177/1362168810375362>
- Swain, M., & Lapkin, S. (2000). Task-based second language learning: The uses of the first language. *Language teaching research*, 4(3), 251–274. <https://doi-org.proxy2.cl.msu.edu/10.1177/13621688000400304>
- Teng, F. (2020). Retention of new words learned incidentally from reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research*, 24(6), 785–812. <https://doi.org/10.1177/1362168819829026>
- Tian, L., & Hennebry, M. (2016). Chinese learners' perceptions towards teachers' language use in lexical explanations: A comparison between Chinese-only and English-only instructions. *System*, 63, 77–88. <https://doi.org/10.1016/j.system.2016.08.005>
- Tian, L., & Jiang, Y. (2021). L2 proficiency pairing, task type and L1 Use: A mixed-methods study on optimal pairing in dyadic task-based peer interaction. *Frontiers in Psychology*, 12, 699774. <https://doi.org/10.3389/fpsyg.2021.699774>
- Tian, L., & Macaro, E. (2012). Comparing the effect of teacher codeswitching with English-only explanations on the vocabulary acquisition of Chinese university students: A lexical focus-on-form study. *Language Teaching Research*, 16(3), 367–391. <https://doi.org/10.1177/1362168812436909>
- Tognini, R., & Oliver, R. (2012). L1 use in primary and secondary foreign language classrooms and its contribution to learning. In E.A. Soler & M. Safont-Jordà (Eds.), *Discourse and language learning across L2 instructional setting* (pp. 53–77). Brill. https://doi.org/10.1163/9789401208598_005
- Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, 69(2), 405–439. <https://doi.org/10.1111/lang.12335>
- Tu, W., & Zhou, X.-H. (1999). A Wald test comparing medical costs based on log-normal distributions with zero valued costs. *Statistics in Medicine*, 18(20), 2749–2761.

- 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), 1–25.
<https://doi.org/10.1080/09588221.2012.692384>
- Turnbull, M. (2001). There is a role for L1 in foreign and second language teaching, but.... *Canadian Modern Language Review*, 57(4), 531–540.
- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69(3), 559–599. <https://doi.org/10.1111/lang.12343>
- van Hell, J. G., & Kroll, J. F. (2012). Using electrophysiological measures to track the mapping of words to concepts in the bilingual brain. *Memory, Language, and Bilingualism*, 126–160. <https://doi.org/10.1017/cbo9781139035279.006>
- VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12(3), 287–301.
<https://doi.org/10.1017/S0272263100009177>
- Varol, B., & Erçetin, G. (2021). Effects of gloss type, gloss position, and working memory capacity on second language comprehension in electronic reading. *Computer Assisted Language Learning*, 34(7), 820–844. <https://doi.org/10.1080/09588221.2019.1643738>
- Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, 61(1), 219–258. <https://doi.org/10.1111/j.1467-9922.2010.00593.x>
- Vraciu, A., & Pladevall-Ballester, E. (2022). L1 use in peer interaction: Exploring time and proficiency pairing effects in primary school EFL. *International Journal of Bilingual Education and Bilingualism*, 25(4), 1433–1450.
<https://doi.org/10.1080/13670050.2020.1767029>
- Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15(2), 130–163.
<https://doi.org/10.1177/003368828501600214>
- Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading: evidence from eye-tracking. *Studies in Second Language Acquisition*, 40(4), 883–906.
<https://doi.org/10.1017/S0272263118000177>
- Watanabe, Y. (2020). Talking to self while writing: Second-language writers' languaging processes and reflections. In W. Suzuki & N. Storch (Eds.). *Languaging in language learning and teaching: A collection of empirical studies* (pp. 197–216). John Benjamins.

- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46–65. <https://doi.org/10.1093/applin/aml048>
- Webb, S. (Ed.). (2019). *The Routledge handbook of vocabulary studies*. Routledge.
- Webb, S., & Chang, A. C. S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, 19(6), 667–686. <https://doi.org/10.1177/1362168814559800>
- Webb, S., & Chang, A. C. S. (2022). How does mode of input affect the incidental learning of collocations? *Studies in Second Language Acquisition*, 44(1), 35–56.
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Xu, J., & Fan, Y. (2021). Task complexity, L2 proficiency and EFL learners' L1 use in task-based peer interaction. *Language Teaching Research*, 136216882110046. <https://doi.org/10.1177/13621688211004633>
- Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading?: A meta-regression analysis. *Studies in Second Language Acquisition*, 42(2), 411–438. <https://doi.org/10.1017/S0272263119000688>
- Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning and Technology*, 10(3), 85–101.
- Yu, S., & Lee, I. (2014). An analysis of Chinese EFL students' use of first and second language in peer feedback of L2 writing. *System*, 47, 28–38. <https://doi.org/10.1016/j.system.2014.08.007>
- Zhang, C., & Ma, R. (2021). The effect of textual glosses on L2 vocabulary acquisition: A meta-analysis. *Language Teaching Research*, 136216882110115. <https://doi.org/10.1177/13621688211011511>
- Zhao, T., & Macaro, E. (2016). What works better for the learning of concrete and abstract words: Teachers' L1 use or L2-only explanations?: Teachers' L1 use or L2-only explanations. *International Journal of Applied Linguistics*, 26(1), 75–98. <https://doi.org/10.1111/ijal.12080>

APPENDIX A: TASK INSTRUCTIONS

A1. Instruction for the Reading Task

下面你将读到一本英文悬疑小说。你可以在手机，平板或者电脑上阅读。阅读小说的目的是理解小说内容。在读的过程中，你将会回答一些关于小说内容的问题。

You will be reading a thriller in English. You can do the reading on your mobile phone, tablet, or laptop. The purpose of the reading is to understand the content. You will be tested on your comprehension during and after reading. You will answer questions regarding the content of the book.

一些词以蓝色和下划线标记。当你点击这些词时，一个窗口会弹出，窗口中有词的定义。如需要了解该词词义，请点击该词。看完词义后，请再次点击，关闭窗口。

这些词可能会在阅读中反复出现，但词义窗口链接只会出现一次。

Some words in the reading are colored in blue and underlined. When you click one of those words, a window will pop up and you will see the definition of the word. You can click those words if you'd like to. **Please click the words again to close the pop-up windows.**

第一天阅读 18 页，第二天阅读 7 页。阅读不限时间。阅读中请不要查阅字典等其他资料。翻页速度较慢时，请不要刷新页面。每页只能阅读一次。翻页后不能再往回读。

You will be reading 18 pages on the first day and 8 pages on the second day. There is no time limit for the reading. Please do not consult the dictionary or any other resource. When the page is loading slowing after you click the 'next page' button, please do not refresh the page. **You can only read each page once.**

A2. Instruction for the Vocabulary Size Test

这是一个词汇量调查。每一道题目的左边是词的定义。请为这些词的定义选择意思相近或对应的词。例如，下图中第一个词义“land with water all around it” 对应词 island, 则在 island 下打钩。每题有 3 个词义，6 个单词，所以有三个单词无对应词义。请尽量不要空题。请独立作答，不要查阅字典或咨询他人。一共 5 页，每页 10 题。限时 30 分钟！

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

This is a test of your vocabulary size. On the left are word definitions. Please match the definitions with corresponding words on the right. For instance, in the image below the first word definition “land with water all around it” corresponds to the word ‘island’. In this case, put a tick under ‘island’. In each question, there are 3 word definitions and 6 words. Therefore, 3 of the words will not match any of the word definitions. Please try to answer all of the questions. Do not consult the dictionary or any other resource. There are 5 pages in the test, with each page containing 10 questions. The time limit is 30 minutes!

A3. Instruction for the Meaning Recall Test

请写出词的意思。可以是词义或者近义词。可以使用中文或英文。不知道答案的，写？，共 24 题。

Please type down the definitions for the words below. You can write down the exact definition or the synonym. You can use Chinese or English. If you don’t know the answer to the question, write ?. 24 items in total.

A4. Instruction for the Meaning Matching Test

这是一个考查词义的配对题。你会看到带编号的 26 个词义和 24 个词。请为词语选择对应的词义编号。

This is a matching test on word knowledge. You will see 26 numbered word definitions and 24 words. Please match the words and their word meanings.

A5. Instruction for the Self-paced Reading Test

这是一个考查句子理解的测试。你会读到一些句子。一开始，屏幕上只有*号。*号表示句子开始的位置。请按**空格键**让测试继续。每按一次**空格键**出现一个词。句子后面会跟有阅读理解题。如觉得题目中信息符合句子内容，用鼠标按 Yes，不符合按 No。

This is a sentence comprehension task. You will read some sentences. At first, you will only see a *. This is where the sentence will start. When you are ready, press **the space bar**. Every time you press **the space bar**, a word of the sentence will appear until the sentence ends. After each sentence, there will be a statement. If you think the statement matches the content the sentence you just read, press the Yes button using your mouse. If not, press the No button.

请按照正常速度阅读句子。一开始是练习。练习结束测试正式开始时会有提示。请按空格键进入测试。

Please read the sentences at your normal pace. You will go through some practice trials at the beginning. When the practice ends, you will be notified. Press the space bar to continue.

A6. Instruction for the Exit Questionnaire

以下是一个关于刚才完成的任务的调查。各个选项无好坏之分，请如实回答。

Next is a survey about the reading task you just finished. Please answer honestly. There is no good or bad answer.

APPENDIX B: VOCABULARY PROFILE OF THE READING MATERIAL

Below lists the percentages of words in the reading material from certain frequency levels (e.g., K1, K2...). K1 represents the most frequent 1000, i.e., 1k, word families, and K2 represents the second 1000, and so on.

K1: 95.9%; K2: 98.3%; K3: 98.7%; K4: 99%; K6: 99.4%; K11: 99.5%

APPENDIX C: TARGET WORD CHARACTERISTICS

Table C1

Target Word Characteristics

Origin	Pseudowords	Part of speech	Frequency	Length (letters)	Orthographic neighbor	Mean bigram frequency	Mean RT
al words							
cop	latpin	Noun	19	6	1	2,789.60	854.20
mafia	valoon	Noun	13	6	1	2,108.00	863
senato	haron	Noun	9	5	3	2,714.75	874.77
r							
trailer	corax	Noun	10	5	2	2,107.75	826.88
murde	caudam	Noun	5	6	1	821.4	804.74
r							
crawl	elile	Verb	4	5	2	2,332.50	871.36
fright	ginge	Verb	7	5	4	3,150.75	856.15
en							
grab	ameld	Verb	5	5	1	1,187.25	805.70
trial	lodice	Noun	3	6	1	1,561.00	893.26
whisp	loupe	Verb	2	5	2	1,303.00	851.52
er							
violen	bitser	Adjective	1	6	1	2,398.40	824.19
t							
shock	appent	Noun	8	6	1	1,975.20	873.17

Table C1 (cont'd)

tail	glunk	Noun	4	5	2	789.75	842.56
pipe							
garag	goncho	Noun	2	6	1	1,815.20	855.53
e							
whisk	vandier	Noun	9	7	1	2,502.33	806.11
ey							
detect	toplin	Noun	5	6	1	2,507.40	854.59
ive							
wild	baggod	Adject	1	6	1	539.2	839.55
		ive					
lorry	jact	Noun	1	4	4	908.333	800.90
innoc	agloat	Adject	1	6	1	1,458.00	854.63
ent		ive					
body	hoag	Noun	27	4	2	799.333	847.50
gun	tonger	Noun	7	6	2	3,452.80	839.07
guard	lancid	Noun	2	6	2	1,746.40	854.42
custod	claft	Noun	5	5	2	747.5	816.23
y							
prison	maive	Noun	8	5	5	1,343.25	803.52
<i>Mean</i>			6.5	5.5	1.83	1794.13	842.23

Table C1 (cont'd)

<i>SD</i>	6.1	.72	1.13	826.40	25.95
	3				
<i>Range</i>	1-	4-7	1-4	593-3453	800-
	27				893

APPENDIX D: GLOSSES

Table D1

Glosses

Target	Pseudowords	L2 glosses	L2 gloss	L1 gloss	L1 gloss
words			length		length
			(words)		(characters)
cop	latpin	police not in uniform	4	便衣警察	4
mafia	valoon	a criminal organization	3	犯罪集团	4
senator	haron	an important politician	3	重要政治 人物	6
trailer	corax	a car used as a house for people to live	8	供人居住 的改装车	8
murder	caudam	a murder that is planned ahead	6	有预谋的 谋杀	6
crawl	elile	move on hands and knees	5	匍匐前进	4
frighten	ginge	make someone nervous	3	使人害怕	4
grab	ameld	to hold suddenly	3	突然抓住	4
trial	lodice	the process to decide if someone committed a crime	9	审判	2
whisper	loupe	speak quietly	2	小声说话	4

Table D1 (cont'd)

violent	Bitser	likely to hurt someone	4	有暴力倾 向的	6
shock	Appent	a unpleasant surprise	3	让人不愉 快的惊喜	8
tail pipe	Glunk	where car smoke gets out	5	汽车排气 管	5
garage	Goncho	a room to store things	5	杂物存储 间	5
whiskey	Vandier	strong alcohol	2	烈酒	2
detective	Toplin	police who gather information about a crime	7	负责搜集 案件信息 的警察	11
wild	Baggod	crazy, out of control	4	疯狂，失 控的	5
lorry	Jact	a truck	2	货车	2
innocent	Agloat	having no knowledge of something	5	不知情的	4
body	Hoag	a dead body	3	死尸	2
gun	Tonger	a small gun	3	小手枪	3

Table D1 (cont'd)

guard	lancid	someone who protects people	4	安保人员	4
custody	claft	being kept in prison	4	被关押	3
Table D1 (cont'd)					
prison	maive	prison for children	3	儿童看守 所	3
<i>Mean</i>			4.17		4.54
<i>SD</i>			1.83		2.17

APPENDIX E: EXIT QUESTIONNAIRE

(English translations do not appear in the actual questionnaire.)

1. 阅读过程中, 你点击查阅词汇的频率大约是 (0%表示没有查阅, 100% 表示查阅了所有词汇)



2. 查阅词汇时, 你的目的是什么? (如, 100%因为理解文章, 或70%因为学习单词)

(1) 理解文章



(2) 学习单词



(3) 如查阅词汇的目的不是理解文章或学习单词, 请注明

3. 有哪些令你印象深刻的帮助理解文章的词汇? (列出词汇所对应的单词即可)

4. 有哪些令你印象深刻的帮助学习单词的词汇? (列出词汇所对应的单词即可)

5. 如果你跳过了一些词汇, 原因是什么? (如, 50%的时候是因为已经猜到词的意思, 或70%的时候因为觉得词的意思不重要) 如果你查阅了所有词汇, 请跳过。

(1) 我已经猜到词的意思了



(2) 我不需要查询词的意思也可以理解文本



(3) 我觉得知道词的意思不重要



(4) 给出的词的意思对我没有帮助



(5) 如以上没有符合你的情况的原因, 请注明

Next

1. How often did you check the word definitions (0% means you didn't check at all, 100% means you checked all of them)

2. What was your purpose when you check the word definitions? (e.g., 100% for reading comprehension, or 70% for word learning)

(1) reading comprehension

(2) word learning

(3) other

3. Are there any word definitions that you found particularly helpful for reading comprehension? (list the corresponding words)

4. Are there any word definitions that you found particularly helpful for word learning? (list the corresponding words)

5. If you skipped some word definitions, what were the reasons? (e.g., 50% of the time, it was because you already guessed the meaning, or 70% of the time because you didn't think the meaning was important). If you checked all the word definitions, please skip this.

(1) I have guessed the word meaning.

(2) I didn't need the word definitions to understand the reading.

(3) I didn't think knowing word meanings was important.

(4) The word definitions were not helpful.

(5) Others.

6.你可以理解给出的词义吗? (100%为全部理解, 0%表示一个词义也不理解)



7.你认为词义对理解文章有帮助吗? (0%表示完全没有帮助; 100%表示非常有帮助)



8.你认为词义对学习单词有帮助吗? (0%表示完全没有帮助; 100%表示非常有帮助)



9.阅读过程中, 你有特意学习单词吗?

- ☐ 有
☐ 无

10.实验过程中, 你有猜到最后会有考查词义的测试吗?

- ☐ 有
☐ 无

你享受阅读的过程吗? (0% 一点也不享受; 100% 非常享受)



Next

6. Can you understand the word definitions? (100% means you understood them all, 0% means you didn't understand any).

7. Do you think the word definitions were helpful for reading comprehension (0% means not helpful at all, 100% means very helpful)

8. Do you think the word definitions were helpful for word learning (0% means not helpful at all, 100% means very helpful)

9. During reading, did you try to memorize words?

Yes

No

10. During reading, did you anticipate word tests afterwards?

Yes

No

Did you enjoy the reading?(0% means not at all, 100% means a lot)

APPENDIX F: LANGUAGE BACKGROUND QUESTIONNAIRE

(English translations do not appear in the actual questionnaire.)

姓名 Name

实验 ID Participant ID

性别 Sex

男

女

不想透露

年龄 Age

请问您在什么年级？ What is your academic status?

大一 Freshman

大二 Sophomore

大三 Junior

大四 Senior

研究生 Masters student

博士生 PhD student

已经毕业 Graduated

请问您有在国外上课的经历吗？ Have you been to school overseas?

有 Yes

无 No

请问您在国外上课的时长为：（例：1 年 3 月） How long have you been to school overseas?

(e.g., 1 year 3 month)

年 year_____

月 month_____

请问您有在国外居住的经历吗？ Have you lived overseas?

有 Yes


无 No

您在国外居住的具体情况是： The details of your stay overseas

	时长（月） Duration (month)	什么时候(例： 2016-2018) When (e.g., 2016-2018)	使用的语言 Language used	居住或停留的 目的 Purpose of stay
国家 1 Country 1				
国家 2 Country 2				
国家 3 Country 3				

请自我评价您的英语水平 (1=差, 10=非常好) Please self-assess your English proficiency (1 = poor, 10 = excellent)

0 1 2 3 4 5 6 7 8 9 10

阅读 reading	
------------	--

写作 writing	
听力 listening	
口语 speaking	
总体 overall	

请提供您完成过的英语测试的成绩 Please tell us scores of standardized English tests you have taken

	总分 score	测试年份（例：2020） year taken (e.g., 2020)
四级 CET4		
六级 CET6		
专四 TEM4		
专八 TEM8		

您几岁开始学习英语？ At what age you start learning English?

您接受了多少年课堂英语教育？ How many years of classroom English learning have you had?

现在，每周您上多少小时的英语课？（没有上英语课请填 0） How many hours of English instruction are you having? (put 0 if you are not taking any English instruction)

您课外学习英语的方式：（可多选） Ways of learning English outside the classroom (you can choose multiple options)

☐

看英语书 Reading English books

☐

看英语电影电视 Watching English movies or TV series

☐

听英语歌 Listening to English songs

☐

用英语对话 Talking to others in English

☐

其他 Others_____

对于您的语言学习背景，请问您还有其他信息想要告诉我吗？

Do you have anything else to say about your English learning?

APPENDIX G: STIMULI FOR THE SELF-PACED READING TEST

Table G1

Stimuli

Sentence (with pseudowords)	Real word	Nonword	Pseudoword meaning
Jason saw a latpin in front of the shop.	police	royate	police not in uniform
They talked about the valoon very often during dinner.	student	remude	a criminal organization
He took a photo for the haron and his wife.	teacher	persh	an important Wpolitician
He designed a corax for his friends.	house	dulpit	a car used as a house for people to live
A caudam took place last year here.	party	pidet	a murder that is planned ahead
He tried to elile while looking at his phone.	walk	arrang	move on hands and knees
I always ginge him in front of people.	kiss	fattice	make someone nervous
He tried to ameld me when I was reading.	touch	premoX	hold suddenly
People are waiting to the lodice to begin at eight.	show	paisin	the process to decide if someone committed a crime
They always loupe to each other.	smile	gatnip	speak quietly

Table G1 (cont'd)

He is a bitser student at school.	helpful	gustre	likely to hurt someone
This reminded him of an appent at the same time last year.	event	feload	an unpleasant surprise
He tried to fix the glunk but he failed.	chair	theath	where car smoke gets out
He read a piece of news about his goncho when he was eating.	company	ostane	a room to store things
He buys vandier every three month.	meat	glunter	strong alcohol
He shook hands with the toplin when they met.	person	esate	police who gather information about a crime
He was baggod after he knew that.	happy	phronic	crazy, out of control
He saw a jact next to his house.	bird	plunt	a truck
He looks agloat all the time.	confused	drimful	having no knowledge of something
He found a hoag next to the tree.	mouse	vertin	a dead body
He took out a tonger all of a sudden.	knife	sluster	a small gun
A lancid is standing in front of the door.	soldier	scrib	someone who protects people
The writer chose claft as the topic.	society	epema	being kept in prison

Table G1 (cont'd)

There is a maive in the north part of the hospital squan prison for children
city.

APPENDIX H: MIXED-EFFECTS MODELLING

Table H1

Mixed Models Specifications

Models	Specifics
Engagement: gloss checking	gloss checking ~ group + vtotal + group * vtotal + (1 participant) + (1+group+zvtotal number)
Engagement: gloss time	gloss time ~ group + vtotal + group * vtotal + (1 participant) + (1+group+zvtotal number)
Matching Immediate	accuracy ~ group + FoO + vtotal + glosstime + group * FoO * vtotal + group * FoO * glosstime + group * vtotal * glosstime + group * glosstime + (1 + FoO + glosstime participant) + (1 + group + glosstime + vtotal number)
Matching Delayed	accuracy ~ group + FoO + vtotal + glosstime + group * FoO * vtotal + group * FoO * glosstime + group * vtotal * glosstime + group * glosstime + (1 + FoO + glosstime participant) + (1 + group + glosstime + vtotal number)
Recall Immediate	accuracy ~ group + FoO + vtotal + glosstime + group * FoO * vtotal + group * FoO * glosstime + group * vtotal * glosstime + group * glosstime + (1 + FoO + glosstime participant) + (1 + group + glosstime + vtotal number)
Recall Delayed	accuracy ~ group + FoO + vtotal + glosstime + group * FoO * vtotal + group * FoO * glosstime + group * vtotal * glosstime + group * glosstime + (1 + FoO + glosstime participant) + (1 + group + glosstime + vtotal number)

Table H1 (cont'd)

Self-paced Immediate	$\begin{aligned} \text{rt} \sim & \text{condition} + \text{FoO} + \text{vtotal} + \text{glosstime} + \text{condition} * \text{FoO} * \text{vtotal} \\ & + \text{condition} * \text{FoO} * \text{glosstime} + \text{condition} * \text{vtotal} * \text{glosstime} + \\ & \text{condition} * \text{glosstime} + (1 + \text{FoO} + \text{glosstime} + \text{condition} \\ & \text{participant}) + (1 + \text{glosstime} + \text{vtotal} \text{number}) \end{aligned}$
Self-paced Delayed	$\begin{aligned} \text{rt} \sim & \text{condition} + \text{FoO} + \text{vtotal} + \text{glosstime} + \text{condition} * \text{FoO} * \text{vtotal} \\ & + \text{condition} * \text{FoO} * \text{glosstime} + \text{condition} * \text{vtotal} * \text{glosstime} + \\ & \text{condition} * \text{glosstime} + (1 + \text{FoO} + \text{glosstime} + \text{condition} \\ & \text{participant}) + (1 + \text{glosstime} + \text{vtotal} \text{number}) \end{aligned}$

Note. Specifics refer to R codes for the models. Vtotal: vocabulary size.

Table H2

Manipulation Check for self-paced reading immediate posttest: Position 0, L1 gloss group

	Fixed effects					Random effects			
						By participant		By item	
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>		<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.24 [6.17, 6.31]	.04	174.57	<.001***		.04	.19	.01	.10
Nonword	-.04 [-.08, .003]	.02	-1.83	.07					
Real word	-.03 [-.08, .01]	.02	-1.46	.14					

Table H3*Manipulation Check for self-paced reading immediate posttest: Position 0, L2 gloss group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>t</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.28 [6.17, 6.31]	.04	139.64	<.001***	.05	.23	.01	.12
Nonword	-.003 [-.08, .003]	.02	-.11	.91				
Real word	-.003 [-.08, .01]	.02	-.13	.90				

Table H4*Manipulation Check for self-paced reading delayed posttest: Position 0, L1 gloss group*

	Fixed effects				Random effects			
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>T</i>	<i>p</i>	By participant		By item	
					<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.09 [6.02, 6.16]	.03	176.53	<.001***	.03	.18	.01	.10
Nonword	-.01 [-.05, .03]	.02	-.57	.57				
Real word	-.04 [-.08, -.001]	.02	-1.99	.05				

Table H5*Manipulation Check for self-paced reading delayed posttest: Position 0, L2 gloss group*

	Fixed effects					Random effects			
						By participant		By item	
	<i>Estimate</i> [95% CI]	<i>SE</i>	<i>T</i>	<i>p</i>		<i>Variance</i>	<i>SD</i>	<i>Variance</i>	<i>SD</i>
Intercept	6.12 [6.04, 6.19]	.04	155.23	<.001***		.04	.19	.01	.10
Nonword	-.03 [-.08, .01]	.02	-1.46	.15					
Real	-.01 [-.06, .03]	.02	-.62	.54					
word									