THE LANGUAGE LEARNING JOURNAL
2019, VOL. 47, NO. 5, 608–624
https://doi.org/10.1080/09571736.2017.1349168

Routledge
Taylor & Francis Group

Incidental L2 lexical acquisition in reading: the role of L2-gloss frequency and learner proficiency

Ting Zhao and Juan Ren

School of Language and Communication Studies, Beijing Jiaotong University, Beijing, People's Republic of China

ABSTRACT

The present study aims to investigate the effects of gloss use, L2-gloss frequency and learner proficiency on incidental L2 lexical acquisition. A total of 163 university students in China were assigned to one of the three reading conditions: no gloss (NG), higher frequency L2 gloss (HFLG) and lower frequency L2 gloss (LFLG). The participants read a text for the purpose of comprehension, and then completed immediate and delayed post-tests for recall and recognition. We then surveyed the participants' views on gloss use and L2-gloss types. The results showed that the glossed groups significantly outperformed the NG group in immediate recall and recognition. The gloss advantage, however, was absent in delayed recall but retained in delayed recognition. The gloss-frequency effect was not observed with higher proficiency participants, but among lower proficiency participants, HFLG was superior to LFLG on immediate recall. The questionnaire results revealed that learners favoured HFLG for the ease of encoding and comprehending. Based on these findings, pedagogical implications are advanced for supporting incidental L2 lexical acquisition within formal educational contexts.

KEYWORDS

L2 gloss; vocabulary learning; retention; word frequency; learner proficiency

Introduction

Lexical researchers have distinguished between two main approaches to L2 vocabulary learning: 'intentional', that is, when learners pay explicit attention to lexical items, and 'incidental', that is, when learners are exposed to lexical items through meaning-focused input (for a review, see Schmitt 2008). Within formal EFL instructional contexts, opportunities for intentional learning are severely limited. A case study into secondary English lessons in China found that only 12.24% of targeted vocabulary words were explicitly taught in class (Tang and Nesi 2003). It is thus reasonable to acknowledge the importance of incidental learning for EFL learners' lexical acquisition and development. But learning words in context (i.e. no gloss (NG), no dictionary look-up) does have several limitations. First, learners may have limited cognitive capacity to grasp the general message of input materials, and simultaneously, to acquire new words. Second, learners may fail to identify or even misidentify the target meaning when contextual clues are not informative for inference (e.g. Laufer 1989; Nassaji 2003). Third, even after learners have inferred a word's meaning correctly, its memory trace may not be sufficiently robust for retrieval.

One solution to overcome these limitations is to gloss new words. A large body of literature has demonstrated the superiority of gloss over NG conditions for L2 lexical acquisition (e.g. Hulstijn, Hollander and Greidanus 1996; Jacobs, Dufon and Hong 1994; Ko 2012; Lv, Yao and Du 2005; Miyasako 2002; Plass et al. 2003; Rouhi and Mohebbi 2012; Watanabe 1997; Yanguas 2009). Some of these studies have found that glosses in the L2 were just as, or even more, effective as L1 glosses

CONTACT Ting Zhao tingzhao@bjtu.edu.cn, tingzhaocn@qq.com

(Jacobs, Dufon and Hong 1994; Ko 2012; Yoshii 2006; but also see Miyasako 2002) and attitude surveys have shown that students, especially higher proficiency learners, favour L2 over L1 glosses (Jacobs, Dufon and Hong 1994; Ko 2012; Miyasako 2002).

Further, the 'frequency-of-occurrence' effect has been well documented in the psycholinguistic literature (e.g. Balota and Chumbley 1984; Barry, Morrison and Ellis 1997; Schilling, Rayner and Chumbley 1998). There are typically two options for glossing a target word (e.g. *swift*) in the L2 if learners have reached an intermediate level of proficiency: the choice is between a higher frequency word (e.g. *quick*) or a lower frequency word[1] (e.g. *rapid*). The present study set out to examine the effectiveness of these two types of L2 glosses for L2 lexical acquisition and retention. The target words fell between the 3K and 8K frequency bands. As suggested by Schmitt and Schmitt (2014), this 'mid-frequency' vocabulary needs more investigation. We also surveyed students' opinions on the use of higher frequency L2 gloss (HFLG) versus LFLG during text reading.

Literature review

The effectiveness of L1 versus L2 glosses

Incidental learning without any input modification does lead to some learning of words but the learning outcomes can by no means be called encouraging. The literature has generally reported about 1 pick-up out of 10 target words (Day, Omura and Hiramatsu 1992; Horst et al. 1998; Hulstijn 1992, exp. 1; Pitts, White and Krashen 1989). The most cited explanation for these ineffective outcomes is Schmidt's (1990, 1994) 'noticing' hypothesis, which postulates that attention is needed for completing any memory task. It is possible that incidental learners do not even notice the presence of new words, let alone memorise them.

One way to facilitate L2 lexical acquisition is to gloss target words in input materials. For example, in Jacobs, Dufon and Hong (1994) research, about 5 out of 32 target words were recalled at immediate post-test in the NG condition, whereas 10 and 9 target words were recalled in the L1- and L2-gloss conditions, respectively. However, the glossing advantage was not retained four weeks later. A more recent study by Ko (2012) found that about 8 out of 16 target words were correctly recognised at immediate post-test in the NG group, whereas about 11 words were recognised in both L1- and L2-gloss conditions. The delayed post-tests confirmed the superiority of gloss over NG conditions.

Given the superiority of gloss over NG conditions, a large body of research has attempted to identify the best way of presenting glosses in terms of their placement (e.g. Holley and King 1971; Jacobs, Dufon and Hong 1994), the frequency of occurrence of the vocabulary targeted (e.g. Hulstijn, Hollander and Greidanus 1996; Rott 1999), the presentation format (Hulstijn 1992; Rott, Williams and Cameron 2002; Watanabe 1997) and the presentation language (e.g. Jacobs, Dufon and Hong 1994; Ko 2012; Miyasako 2002; Yoshii 2006). The studies most relevant to the present one are those comparing the effectiveness of L2 and L1 glosses. Jacobs, Dufon and Hong (1994), Yoshii (2006) and Ko (2012) found that L1 and L2 glosses were equally beneficial for L2 lexical acquisition. Jacobs, Dufon and Hong's research, for example, involved 85 native speakers of English enrolled in a Spanish language programme. They were assigned to one of the three conditions: NG, L2 (Spanish) glosses and L1 (English) glosses. After reading a passage, they were asked to recall the passage and to translate target words from Spanish to English. The results showed that their passage recall did not vary significantly with the learning conditions. In the immediate and delayed post-tests, the effectiveness of L2 and L1 glosses was comparable. Over half the participants (53%) favoured L2 glosses as long as these glosses were comprehensible. Jacobs, Dufon and Hong concluded that L2 educators should try to use L2 glosses as much as possible.

Miyasako (2002) considered learners' L2 proficiency in examining the language-of-gloss effect. The participants were 187 native speakers of Japanese at a senior high school. They were randomly allocated to one of the six conditions: multiple-choice glosses (L1 versus L2), standard glosses (L1 versus L2), NG and no reading. A vocabulary test on 15 target words was administered immediately after the

intervention and 18 days later. The results showed that L2 glosses were more effective than L1 glosses at immediate post-test; however, this advantage was not retained at delayed post-test. A significant interaction was found between L2 proficiency and language-of-gloss in that L2 glosses were more effective for higher proficiency learners' immediate recall, whereas L1 glosses were more effective for lower proficiency learners' immediate recall.

Previous research on the language-of-gloss effect has thus demonstrated the effectiveness of L2 glosses for lexical acquisition and retention. This is one of the reasons why we believe L2 glosses deserve further examination as a pedagogic tool. The provision of L2 glosses is likely to increase learners' exposure to the target language, and reduce activation of the L1. Practically, in multilingual classes where students have different mother tongues, teachers have no alternatives but to rely on the target language. Additionally, teaching materials designed by native speakers usually make use of L2 exclusivity.

The gloss-frequency factor

Using the British National Corpus (BNC), we counted the frequency of target words and that of L2 glosses, respectively, in the four studies summarised in Table 1. The frequency of most target words ranged between 3K and 8K, with the remainder from a mixture of higher or lower frequency bands. A large proportion of L2 glosses fell within the 1K band: about three-fourths in Jacobs, Dufon and Hong (1994) and about three-fifths in Ko (2012). Such heavy use of 1K band glosses appears to be common in current L2 classrooms, which is understandable given that teachers may be concerned about the negative outcomes of vocabulary learning via LFLGs. However, as Schmitt and Schmitt (2014) pointed out, if teacher talk and input materials is dominated by high-frequency words, there must be a serious lack of repetition of low- and mid-frequency words. Such pedagogical practice may pose a challenge to learners' vocabulary expansion. Given this theory-practice divide, it is necessary to identify whether LFLGs were less effective than their higher frequency L2 counterparts in terms of lexical acquisition.

Advocates for HFLG and LFLG have their own theoretical arguments. Psycholinguistic research on monolinguals has shown that accessing high-frequency words takes considerably less effort than accessing low-frequency words (e.g. Balota and Chumbley 1984; Barry et al. 1997; Forster and Chambers 1973; Oldfield and Wingfield 1965). The most likely explanation draws on the episodic model (Goldinger 1998; Hintzman 1986) where input experiences are responsible for memory consolidation. Higher frequency words appear to leave stronger memory traces in the brain than lower frequency words, thus inspiring the development of 'abstract' echoes. Thus, it is argued that HFLGs can be accessed more easily than LFLGs, leaving more cognitive capacity available for connecting the new words to the HFLGs.

Conversely, several arguments appear to support the use of LFLGs. According to Craik and Lockhart's (1972) depth of processing theory, deep-level processing involves more mental effort than

Table 1. Frequency counts of target words and L2 glosses in four studies.

Study	Target words	L2 glosses
Jacobs, Dufon and Hong (1994)	$N = 16$ 3K–8K: 87.5%; 9K–11K: 12.5%	$N = 26$ 1K: 76.92%; 2K: 19.23%; 5K: 3.85%
Miyasako (2002)	$N = 15$ 2K: 13.33%; 3K–8K: 86.67%	Not available for this analysis
Yoshii (2006)	$N = 14$[a] 1K–2K: 28.57%; 3K–6K: 71.43%	Not available for this analysis
Ko (2012)	$N = 16$ 1K–2K: 11.76%; 3K–8K: 56.25%; 9K–13K: 31.25%	$N = 30$ 1K: 59.38%; 2K: 21.88%; 3K: 9.38%; 4K: 9.38%

[a]In Yoshii's (2006) study, the text consisted of 14 target words and 6 distractors. The author provided the text in the appendix but did not state clearly which words were target words. We thus excluded the six higher frequency words as distractors from this analysis.

surface-level processing. Greater depths of processing should lead to better memorisation. Processing LFLGs arguably requires more mental effort than processing higher-frequency glosses, and thus lower-frequency glossing should yield better vocabulary retention. Glanzer and Adams (1990), for example, found a 'mirror effect' for word frequency, that is, higher hit rates for low-frequency as opposed to HFLGs, and conclude that low-frequency glosses draw more attention. LFLGs may be attended to, and then encoded, more – or at least equally – effectively than HFLGs.

Up to the present, there has been a lack of empirical evidence on how L2-gloss frequency affects incidental L2 lexical acquisition. Advocates for HFLG and LFLG both have their respective theoretical justifications. Under such circumstances, we tentatively hypothesise that the effectiveness of LFLG and HFLG may not be significantly different.

L2 lexical acquisition and the bilingual lexicon

In the case of successive bilinguals, L1 lexicons and their corresponding concepts have already been developed to a great extent prior to L2 lexical acquisition. Various models have been proposed to explain how L1 and L2 lexicons are structured in bilinguals. For example, in the Revised Hierarchical model (RHM, Kroll and Stewart 1994, see Figure 1) it is assumed that L1 and L2 words share a single conceptual system: at the outset, L1-concept connections are stronger than L2-concept connections, and L2 words need to access their L1 counterparts before accessing the conceptual system. With increasing L2 proficiency, bilinguals become less reliant on the L2–L1 routes but more on the L2-concept routes (the developmental hypothesis, Kroll and De Groot 1997). In other words, the mechanisms underlying L2 lexical processing are actually modulated by learner proficiency.

Another model worth mentioning is Jiang's (2000, 2004) three-stage model, which specifies the developmental trajectory of individual L2 lexicons. L2 lexical acquisition proceeds consecutively through an L2–L1 word association stage, an L2 lemma (i.e. semantic and syntactic representations) mediation stage and a full L2 integration stage. At the very beginning, an L1 word mediates the relationship between an L2 word and its concept. As learners' contextualised exposure to the L2 word increases, the magnitude of the conceptual route will outweigh that of the lexical route.

What do the above two models of bilingual lexicons imply for the present study? First, the mental representation of an L2 word is sensitive to a learner's L2 proficiency. According to the developmental hypothesis (Kroll and De Groot 1997), lower proficiency participants are assumed to rely more on lexical associations than higher proficiency learners. Second, considering Jiang's model, participants' L1 knowledge is assumed to sustain the initial stage of L2 lexical acquisition. Learners needed to be continuously exposed to a target word in order to complete the third stage of full L2 integration.

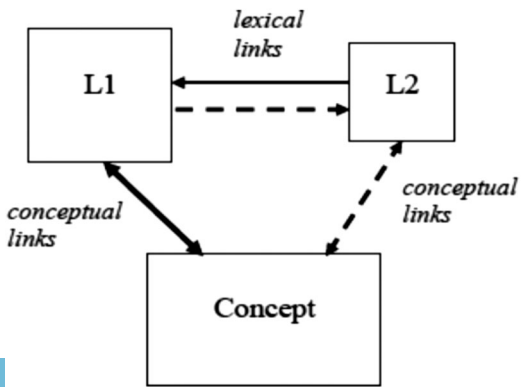

Figure 1. Revised hierarchical model. Adapted from Kroll and Stewart (1994).

In summary, the present study sets out to compare the effectiveness of three different learning conditions (NG, HFLGs and LFLGs) for L2 lexical acquisition and retention with higher versus lower proficiency learners. Based on the literature reviewed above, we assume that the NG condition will be less effective than the glossed conditions. However, we cannot identify conclusive evidence or theoretical arguments for the greater effectiveness of either HFLG or LFLG. We, therefore, tentatively hypothesise that these two types of glosses will be equally effective for incidental lexical acquisition. We further assume that learner proficiency will affect word-retention outcomes.

Research questions

(1) Do learners benefit more from glossed conditions than from the NG condition in terms of their incidental vocabulary acquisition and retention?
(2) Do the effects of HFLG and LFLG differ in incidental vocabulary learning?
(3) Does the factor of learner proficiency (higher versus lower) modulate the effects of NG, HFLG and LFLG on lexical acquisition and retention?
(4) What are learners' views on the use of HFLG and LFLG?

The study

Participants

A total of 175 first-year undergraduate EFL learners (aged 18–19) from six parallel intact classes participated in our experiment. They were native speakers of Mandarin and most had been learning English as a compulsory subject for about 10 years at the time of our data collection (3 years at primary school, 3 years at middle school, 3 years at high school and 1 year at university). Each week participants received six 50-minute sessions of College English training. Each class was randomly assigned to one of the experimental conditions: 58 students to the NG condition, 57 to the HFLG condition and 60 to the LFLG condition.

Three weeks before our experiment, participants took an academic version of the IELTS (International English Language Testing System) test[2] to check their overall proficiency, and a VLT (Vocabulary Levels Test) to measure their vocabulary knowledge. The VLT is an adapted version of the size test (Nation and Belgar 2007). Given the participants' English proficiency, the selected test words (N = 50) were from the first five frequency levels. For each test word, participants needed to select a correct translation from the four options provided.

Based on their IELTS bands and VLT scores, the participants were classified as higher proficiency (IELTS $\geq$ 4.5 and VLT > average score 34.04) and lower proficiency (IELTS < 4.5 and VLT < 34.04). We considered both IELTS and VLT test scores to define the participants' proficiency for cross-validation purposes. We screened out 12 participants, as their IELTS bands were greater than or equal to 4.5 but VLT scores were lower than 34.04, or their IELTS bands were less than 4.5 but VLT scores were higher than 34.04. The final sample size was 163, with 53 participants in the NG condition, 55 participants in the HFLG condition and 55 participants in the LFLG condition. Table 2 displays the descriptive statistics of higher and lower proficiency participants' test scores in each learning condition.

Prior to the conduct of our experiment, we checked baseline comparability between the learning conditions. The normality and homogeneity of variance were tested with Kolmogorov–Smirnov and Homogeneity of Variances tests. Inspection of the data showed that some of the variables were not normally distributed (Kolmogorov–Smirnov test: ps < .5). Nevertheless, ANOVA is believed to be fairly robust against the violation of normality when sample sizes are equal (Field 2013). Given our six nearly equal-sized groups, the baseline data were submitted to one-way ANOVA analysis. No significant differences were found between the learning conditions within the higher proficiency group in terms of IELTS bands, $F(2, 79) = .87$, $p = .42$, and vocabulary size, $F(2, 79) = .19$, $p = .83$; and within the

Table 2. Higher and lower proficiency participants' L2 knowledge in each learning condition.

			Vocabulary knowledge (maximum score = 50)			Overall IELTS band score (maximum score = 9)		
Proficiency levels	Gloss conditions	*n*	*M*	SD	Min.–Max.	*M*	SD	Min.–Max.
Higher proficiency	NG	27	38.11	2.72	35–45	4.83	0.37	4.5–5.5
	HFLG	28	37.96	2.12	35–44	4.71	0.29	4.5–5.5
	LFLG	27	37.67	3.22	35–46	4.78	0.35	4.5–5.5
Lower proficiency	NG	26	29.12	3.20	24–34	3.81	0.49	2.0–4.0
	HFLG	27	30.67	2.87	24–34	3.74	0.68	1.0–4.0
	LFLG	28	30.57	3.16	23–34	3.75	0.63	2.0–4.0

Note: NG: no gloss; HFLG: higher frequency L2 gloss; LFLG: lower frequency L2 gloss.

lower proficiency group in terms of IELTS bands, $F(2, 78) = .10$, $p = .91$, and vocabulary size, $F(2, 78) = 2.11$, $p = .13$. Furthermore, the random allocation of intact classes to one of the learning conditions corroborated baseline comparability.

Reading materials and target words

Two local experienced EFL teachers, each with more than five years of English teaching experience at university, selected a candidate article for our research. The 921-word text was then piloted with 12 students comparable in terms of English learning background and proficiency level to those participating in the main experiment. They were instructed to underline any unknown words and then to indicate whether the provided LFLGs were known or unknown. Based on the results of pilot study, the text was modified to meet the criterion that a reader needs to know about 95% of a text's vocabulary for adequate comprehension (Laufer 1997; Nation 2001). We submitted all words (except pronouns) in the text to the Compleat Lexical Tutor website (Cobb 2000). The lexical profile showed that 96% of the tokens (i.e. running words) were below the 3000-word level, indicating that the text was appropriate for the participants.

We selected 15 words as candidate targets (4 nouns, 8 adjectives, 3 verbs). These items appeared only once in the text and were unknown to the pilot participants. Both HFLG and LFLG were then made available for these items. One week before the learning session, we administered a pre-test in which the 163 participants were instructed to underline any unknown items in a list of 65 words (the 15 candidate target words + 50 course-book words as distractors). The following 12 target words (2 nouns, 7 adjectives, 3 verbs) were unknown to all 163 participants and were selected for our research: *principal*, *converge*, *neglect*, *abiding*, *peril*, *assent*, *prevailing*, *routine*, *unprecedented*, *elated*, *abatement*, *swift*. According to the estimates of BNC, the average frequency of the selected target words was 4.92K ($SD = 1.93$K, range: 3K–8K); and the average length was 7.67 letters ($SD = 2.31$, range: 5–13). The target words were not cognates in L1 translations.

L2 glosses

The target words were glossed using L2 synonyms or L2 explanations. The glosses were chosen based on the participating teachers' knowledge of the students' vocabulary size. The word frequency of L2 synonyms, or the keywords in L2 explanations, was different depending on the learning condition (see Appendix). The glosses appeared in the margin next to the line containing the corresponding target word, as in the following example:

… we live in an **elated**[10] age of transformation **10. elated: very happy**

In order to estimate the participants' pre-knowledge of the words used in L2 glosses, 35 first-year undergraduates not participating in the experiment were asked to provide correct translations for the 24 glosses. Descriptive statistics of the lexical characteristics of the final L2 glosses are given in Table 3. Independent samples *t*-tests showed that the words used in the HFLGs were indeed used

Table 3. Summary of lexical characteristics of HFLG versus LFLG.

	Word frequency			Word length		
Gloss conditions	*M*	*SD*	Min.–Max.	*M*	*SD*	Min.–Max.
Higher frequency ($n = 12$)	1.08	0.29	1.00–2.00	5.42	1.38	4.00–9.00
Lower frequency ($n = 12$)	2.25	0.45	2.00–3.00	5.92	1.44	4.00–9.00

far more frequently than those in the LFLGs, $t(22) = -7.53$, $p < .001$. However, the word length of the HFLGs was not significantly different from that of the LFLGs, $t(22) = -0.87$, $p = .40$.

Assessment instruments

A recall test and a recognition test were administered to the participants to assess their acquisition and retention of the meaning of target items. An example of the recall test is given below:

Please use L1 or L2 to explain the meaning of the following words.

请用中文或者英文解释下列单词的意思。

Principal _____

The recall test was scored as follows: each correct answer (e.g. *principal* = *chief*) received two points, a semantically related answer (e.g. *principal* = *valuable*) received one point, and an incorrect answer was scored zero (e.g. *principal* = *urgent*). The maximum score was 24. Following the recall test, participants also took a multiple-choice recognition test. An extract from this test is given below:

Please choose the best answer from the four translations provided.

请从以下四个选项中选出最佳答案。

Principal

A. 急迫的 (urgent)

B. 主要的 (chief)

C. 严厉的 (strict)

D. 高尚的 (noble)

For the recognition test, a correct choice was awarded two points, and any incorrect choice was scored zero. Both the recall and the recognition tests were repeated after a delay of one week. In order to alleviate the possible Hawthorne effect, 18 course-book words were also included in each immediate and delayed post-test.

Teachers and experimental procedures

The two teachers who participated in our experiment had more than five years of EFL teaching experience. They were comparable in terms of their teaching expertise and pedagogical style. Such comparability was verified by reviewing the teachers' profiles, observing their classes, and interviewing them in person. In order to maximise treatment fidelity, they complied with a pre-determined teaching plan throughout the experiment.

The teaching procedure consisted of a reading phase, an exercise phase and a post-test phase. After the teacher had distributed the reading text, the participants were instructed to read it silently during 10 minutes and to summarise each paragraph in one sentence using the L1. The participants in the NG condition did not receive any treatments related to the L2 target words, while the participants in the HFLG and LFLG conditions were provided with L2 glosses. Once the students had

completed the reading task, the teacher collected the reading materials. This was immediately followed by the two post-tests (i.e. meaning recall and meaning recognition) administered one after the other. To assess the participants' retention of the target items, the two tests were repeated a week later.

The whole study took place over a five-week period. The baseline tests were administered in the first week when we also piloted the appropriateness of the reading material, target words and L2 glosses. We then modified the research instruments where required. In the second week, we interviewed some pilot participants to inform design of the participant questionnaire. In the third week, the participants completed a pre-test without prior notice. In the fourth week, the two post-tests were administered immediately after the intervention, and in the fifth week, the two tests were again completed as delayed post-tests. A questionnaire was also distributed to the participants to seek their opinions on the use of HFLG and LFLG during reading.

L2-gloss preference questionnaire

In order to devise questionnaire items, one-to-one interviews were conducted with 10 pilot participants in the second week, focusing on their attitudes towards HFLG and LFLG. The length of interviews was between 20 and 40 minutes. We transcribed all interviews, and coded and then analysed them. Based on the interview findings, we developed and distributed a preliminary questionnaire to 15 pilot participants. Any unclear or ambiguous items were then modified. The final version of the questionnaire consisted of two sections: (1) ranking the preference order of L2-gloss types and (2) indicating level of agreement with 10 opinion statements on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The 163 participants completed the questionnaire statements in class after completing the delayed post-tests.

Results

Results of recall tests

Table 4 presents descriptive statistics of higher and lower proficiency L2 participants' meaning recall at immediate and delayed post-test. The gloss groups significantly outperformed the NG group at immediate post-test. There was, however, a significant decrease in recall scores for all treatment groups between immediate and delayed testing.

We submitted the data to repeated-measures ANOVA to examine the differences between the six groups. The results showed significant differences between the immediate and delayed post-tests: $F(1, 157) = 363.96$, $p < .001$, $\eta_p^2 = .70$; between learning conditions: $F(2, 157) = 34.59$, $p < .001$, $\eta_p^2 = .31$; and between the two levels of L2 proficiency: $F(1, 157) = 18.94$, $p < .001$, $\eta_p^2 = .11$. Given a significant

Table 4. Descriptive statistics of meaning recall in each learning condition within the two proficiency groups.

		Recall immediate			Recall delayed		
Proficiency	Condition	*M*	*SD*	%	*M*	*SD*	%
Higher (*n* = 82)	NG (*n* = 27)	2.30	0.72	9.58	2.19	0.68	9.13
	HFLG (*n* = 28)	6.18	2.16	25.75	2.54	1.37	10.58
	LFLG (*n* = 27)	6.70	2.30	27.92	2.52	1.28	10.50
Lower (*n* = 81)	NG (*n* = 26)	2.00	0.08	8.33	1.88	0.65	7.83
	HFLG (*n* = 27)	5.67	2.20	23.63	2.19	1.39	9.13
	LFLG (*n* = 28)	3.29	1.44	13.71	1.96	0.92	8.17

Notes: NG: no gloss; HFLG: higher frequency L2 gloss; LFLG: lower frequency L2 gloss. The maximum score was 24.

Table 5. *Post hoc* analysis of meaning recall: effects of gloss use, learner proficiency and time of test.

		Condition (I)	Condition (J)	Mean difference (I–J)	Std. error	Sig.	Mean difference (I–J)	Std. error	Sig.
				Recall immediate			Recall delayed		
Gloss effect	HP	HFLG	LFLG	−0.52	.49	.29	0.02	.30	.95
		HFLG	NG	3.88[a]	.49	.00	0.35	.30	.24
		LFLG	NG	4.41[a]	.50	.00	0.33	.30	.27
		HFLG	LFLG	2.38[a]	.49	.00	0.22	.30	.46
	LP	HFLG	NG	3.67[a]	.50	.00	0.30	.30	.32
		LFLG	NG	1.29[a]	.50	.01	0.08	.30	.79
Proficiency effect	HFLG	HP	LP	0.51	.49	.30	0.35	.30	.24
	LFLG	HP	LP	3.42[a]	.49	.00	0.55	.30	.06
	NG	HP	LP	0.30	.50	.56	0.30	.30	.32
				HP			LP		
Time of test effect	HFLG	Immediate	Delayed	3.64[a]	.27	.00	3.48[a]	.27	.00
	LFLG	Immediate	Delayed	4.19[a]	.28	.00	1.32[a]	.27	.00
	NG	Immediate	Delayed	0.11	.28	.69	0.12	.28	.68

Note: NG: no gloss; HFLG: higher frequency L2 gloss; LFLG: lower frequency L2 gloss; HP: higher proficiency; LP: lower proficiency.
[a]The mean difference is significant at the .05 level.

interaction found between test time, learning condition and learner proficiency: $F(2, 157) = 17.21$, $p < .01$, $\eta_p^2 = .18$, Bonferroni-corrected *post hoc* analyses were conducted to investigate the interaction effects.

First, as shown in Table 5, the effectiveness of HFLG and LFLG was found to be comparable on immediate recall among the higher proficiency learners ($p = .29$), whereas for lower proficiency learners, HFLG was more effective than LFLG ($p < .001$). Thus, gloss frequency was found to modulate lower proficiency participants' immediate recall. Students in the gloss conditions significantly outperformed those in the NG condition ($ps < .001$) on immediate recall. However, the results of the delayed recall test showed no significant differences between the three learning conditions ($ps > 2.3$), indicating that the gloss advantage was not retained for long-term memory.

Second, regarding the proficiency effect, there were no significant differences in word recall between higher and lower proficiency learners in the HFLG condition ($ps > .05$). In the LFLG condition, however, the higher proficiency group (mean score = 6.70) outperformed the lower proficiency group (mean score = 3.29) at immediate post-test ($p < .001$). Thus, the 'proficiency' effect was present in the LFLG condition but absent in the HFLG condition.

Third, in the two glossed conditions, the mean scores of delayed recall were significantly lower than those of immediate recall. Such retention loss was observed regardless of learners' L2 proficiency. It can be concluded that L2 lexical acquisition via L2 glosses was vulnerable to memory loss over time. In the NG condition, there were no significant differences between the two assessment points.

Results of recognition tests

Table 6 shows descriptive statistics of the participants' meaning recognition scores at the two test points. The participants, in general, performed better on word recognition than recall, which is understandable given that recognition is based on identifying previously learned words with external cues, whereas recall proceeds without such cues. The statistical analyses were the same as in the recall analyses. The results showed significant differences between the immediate and delayed post-tests: $F(1, 157) = 312.77$, $p < .001$, $\eta_p^2 = .67$; between learning conditions: $F(2, 157) = 112.29$, $p < .001$, $\eta_p^2 = .59$; and between the two learner proficiency groups: $F(1, 157) = 3.33$, $p < .001$, $\eta_p^2 = .02$. Given a significant interaction found between test time and learning conditions: $F(2, 157) = 88.86$, $p < .001$, $\eta_p^2 = .43$, Bonferroni-corrected *post hoc* analyses were performed to investigate the interaction effect.

First, as shown in Table 7, the gloss conditions were more effective than the NG condition for both immediate and delayed recognition ($ps < .001$). The 'gloss' effect was thus retained in delayed

Table 6. Descriptive statistics of meaning recognition in each learning condition within the two proficiency groups.

		Recognition immediate			Recognition delayed		
Proficiency	Condition	*M*	*SD*	%	*M*	*SD*	%
Higher (*n* = 82)	NG (*n* = 27)	4.37	0.97	18.21	4.15	0.95	17.29
	HFLG (*n* = 28)	10.86	2.58	45.25	7.43	2.10	30.96
	LFLG (*n* = 27)	11.33	3.04	47.21	7.70	2.87	32.08
Lower (*n* = 81)	NG (*n* = 26)	3.69	1.09	15.38	3.31	1.12	13.79
	HFLG (*n* = 27)	10.67	3.25	44.46	7.26	2.36	30.25
	LFLG (*n* = 28)	10.36	2.78	43.17	7.00	2.58	29.17

Notes: NG: no gloss; HFLG: higher frequency L2 gloss; LFLG: lower frequency L2 gloss. The maximum score was 24.

Table 7. *Post hoc* analysis of meaning recognition: effects of gloss use and time of test.

		Condition (I)	Condition (J)	Mean difference (I–J)	Std. error	Sig.	Mean difference (I–J)	Std. error	Sig.
				Recognition immediate			Recognition delayed		
Gloss effect	HP	HFLG	LFLG	−0.48	.64	.46	−0.28	.58	.63
		HFLG	NG	6.49[a]	.64	.00	3.28[a]	.58	.00
		LFLG	NG	6.96[a]	.64	.00	3.56[a]	.58	.00
	LP	HFLG	LFLG	0.31	.64	.63	0.26	.58	.65
		HFLG	NG	6.97[a]	.65	.00	3.95[a]	.59	.00
		LFLG	NG	7.00[a]	.65	.00	3.69[a]	.58	.00
				HP			LP		
Time of test effect	HFLG	Immediate	Delayed	3.43[a]	.33	.00	3.41[a]	.33	.00
	LFLG	Immediate	Delayed	3.63[a]	.33	.00	3.36[a]	.33	.00
	NG	Immediate	Delayed	0.22	.33	.51	0.39	.34	.52

Note: NG: no gloss; HFLG: higher frequency L2 gloss; LFLG: lower frequency L2 gloss; HP: higher proficiency; LP: lower proficiency.
[a]The mean difference is significant at the .05 level.

recognition but not in delayed recall, indicating that lexical measurements affect the 'gloss' effect over time. Among higher proficiency participants, there were no significant differences in either immediate or delayed recognition between the HFLG and LFLG conditions (*ps* > .45). Among lower proficiency learners, there were also no significant differences for recognition between the HFLG and LFLG conditions ($p = .63$). This contrasts with the result noted above for immediate recall where HFLG was found to be more effective than LFLG for lower proficiency learners.

Second, there was a noticeable drop between immediate and delayed recognition in the four gloss groups (*ps* < .001) whereas the NG group did not experience such memory loss over time (*ps* > .51). The participants' performance on recognition and recall tests showed the same pattern of the 'time-of-test' effect in each learning condition.

To sum up, the present study showed that L2 glosses facilitated incidental vocabulary acquisition at immediate post-test. Both higher and lower proficiency learners in the two gloss conditions performed equally well on immediate recognition. The 'proficiency' effect was observed in lower proficiency learners' immediate recall. The 'time-of-test' effect varied with learning conditions: both higher and lower proficiency participants recalled and recognised fewer words at delayed post-test in all four gloss conditions, whereas such memory loss was not observed with the NG condition.

Attitude questionnaire results

The results of the questionnaire (see Table 8) showed that 72.4% of the participants ranked HFLGs as their first option, whereas only 13.5% preferred LFLG to HFLG. The remainder (14.11%) regarded both

Table 8. Student preferences for L2-gloss types.

Ranked preference			
H–L	L–H	B–H–L	B–L–H
72.40%	13.50%	4.91%	9.20%

Note: H: higher frequency L2 glosses; L: lower frequency L2 glosses; B: both types of L2 glosses.

Table 9. Results of questionnaire items: opinions on HFLG versus LFLG.

	Likert scale		Strongly disagree		Neutral		Strongly agree
Statement	*M*	*SD*	1	2	3	4	5
1. The higher the frequency of L2 glosses, the more likely you will arrive at a correct understanding of target words	4.45	0.67	0.00	1.23	6.13	39.26	53.37
2. The higher the frequency of L2 glosses, the less time you will need to react to target words	4.43	0.70	0.00	1.84	6.75	38.04	53.37
3. The higher the frequency of L2 glosses, the more likely you will associate target words with other known words	4.35	0.73	0.00	3.07	6.13	43.56	47.24
4. The higher the frequency of L2 glosses, the stronger your mental images elicited for target words	4.26	0.73	0.00	1.23	12.88	44.79	41.10
5. The lower the frequency of L2 glosses, the more likely you will look up the meaning of target words in a bilingual dictionary	3.89	1.02	1.84	9.20	19.02	38.04	31.90
6. The lower the frequency of L2 glosses, the more likely you memorise the meaning of target words via L1 equivalents	3.70	1.22	7.36	9.20	21.47	30.06	31.90
7. The lower the frequency of L2 glosses, the more likely you will forget the meaning of target words	3.56	0.99	0.00	19.63	22.09	41.10	17.18
8. The lower the frequency of L2 glosses, the less you will be able to memorise the meaning of target words	3.43	1.10	4.91	19.63	16.56	45.40	13.50
9. The lower the frequency of L2 glosses, the less clear you will be about the grammatical class of target words	3.40	1.00	1.84	19.63	28.22	37.42	12.88
10. The lower the frequency of L2 glosses, the less you will be able to use target words in practice	3.37	1.08	6.75	14.72	24.54	42.33	11.66

Notes: 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree. Items are arranged according to the mean ratings in descending order.

types of L2 glosses acceptable. Previous literature (Jacobs, Dufon and Hong 1994; Ko 2012; Miyasako 2002) has assumed that learners' L2 proficiency is a crucial factor in influencing learner attitudes towards gloss types, an assumption we examined here. The results, however, showed that the participants' gloss preferences did not correlate significantly with either their IELTS scores ($p = .41$) or their vocabulary size ($p = .17$). In the present study L2 proficiency did not significantly affect learners' preference for a certain type of L2 gloss.

Regarding the participants' specific reasons for their gloss preferences (see Table 9), about 90% of the participants held that the use of HFLGs facilitated comprehension of target words (Item 1; $M = 4.45$), and shortened the time spent on understanding target words (Item 2; $M = 4.43$). The other reasons that participants agreed upon were: better associations between a target word and other known words (Item 3; $M = 4.35$), stronger mental images aroused for target words (Item 4; $M = 4.26$), and fewer dictionary consultations are involved (Item 5; $M = 3.89$). Overall, the preference for HFLGs is more associated with the early phase of processing L2 glosses (i.e. encoding and comprehension) rather than its later phase (i.e. memorisation and practice).

Discussion

Comparison between glossed and NG conditions

The present study suggests that L2 glosses are beneficial for incidental L2 lexical acquisition. On immediate testing, participants who read an L2-glossed text recalled and recognised more words than those who read a text without L2 glosses. On delayed testing, the gloss advantage disappeared

for recall but was retained for recognition regardless of learners' L2 proficiency. These results are consistent with Jacobs, Dufon and Hong (1994) and Ko's (2012), who compared the effectiveness of glossed conditions with that of an NG condition over time.

Why do the participants recall and recognise more words via a glossed text? First, inferring a word's meaning from its context appears to be somewhat unreliable and ineffective. This type of lexical learning involves explicit or implicit reasoning (Nöth 2014), an opportunistic process during which the quality of lexical inferences is affected by the contextual clues' informativeness and unambiguousness. There are several specific reasons why learners may make incorrect and inaccurate inferences: (1) contextual information is not sufficient for inference (Huckin and Bloch 1993); (2) the semantic boundary between some words is not adequately clear-cut, in which case learners may misidentify the target meaning (e.g. *kayak – canoe*); (3) the inferred meaning either represents a subset of the target meaning (e.g. inferred meaning: *flower*; target meaning: *flora*), or extends far beyond it (e.g. inferred meaning: *animal*; target meaning: *mammal*); and (4) lexical inference draws on word-shape familiarity rather than on contextual information (e.g. *tint – tiny*, Huckin and Bloch 1993).

Another explanation is that unattended information tends to be memorised by chance. Attention is selective per se in that processing information is limited to one's cognitive capacity. Readers may not attend to the presence of target words until their text comprehension is hampered. The provision of glosses does attract learners' attention to the targeted vocabulary. As such, the likelihood of incidental lexical acquisition increases through reading a glossed text. Put simply, noticing is a precondition for intake to occur in language learning (Schmidt 1993, 2001). Noticing enhances learners' chances of encoding L2 glosses, and then makes lexical acquisition more possible.

Even if learners are able to make correct inferences about an unknown word, it has been shown that form-focused, but not meaning-focused, inferences were correlated with lexical retention (Hu and Nassaji 2012). That is to say, the depth with which learners make meaning-focused inferences may not be robust enough for later retrieval. Additionally, ease of inference may lead to lower chances of word retention (Hu and Nassaji 2012; Pulido 2009), probably because a less demanding inference involves a lower level of learning load.

We found that L2 proficiency did not affect learners' ability to benefit from reading an NG text. This result is not surprising, as reading a text alone is not highly effective for L2 lexical acquisition. Hence, the NG group manifested a 'floor effect', where participants, regardless of L2 proficiency, acquired a very limited number of target words.

Why does the gloss advantage disappear in delayed recall but persist in delayed recognition? Vocabulary assessment tools may be a confounding variable in gauging learners' lexical acquisition and retention. Meaning recall is undoubtedly more demanding than recognition because the former lacks cues to facilitate retrieval. Recall memory without any further consolidation may deteriorate simply with the passage of time. Hence, all participants, regardless of their L2 proficiency and learning conditions, could only recall about one target word at delayed post-test. By contrast, word recognition in a multiple-choice format is far less demanding; this might explain why the gloss advantage was found to be retained for delayed recognition.

Comparison between HFLG and LFLG conditions

The present study showed that there were no significant differences between the HFLG and LFLG conditions in terms of higher proficiency participants' immediate recall and recognition. It is possible that the 'gloss-frequency' effect was reduced as learners had already achieved a certain level of proficiency. According to the results of baseline tests, the higher proficiency participants were classified as upper-intermediate learners. This explanation is consistent with Ko (2012) view that understanding L2 glosses correctly requires at least an intermediate level of L2 proficiency. We performed an analysis on the percentage of L1 versus L2 use in the test answers given by the participants. Higher proficiency participants were found to use the L1 to recall a majority of the target words (approximately

89%). Such heavy reliance on L1 knowledge is compatible with Jiang's (2000, 2004) three-stage model, where the L2–L1 route underlies the processing of newly learned words.

Lower proficiency participants recalled fewer words in the LFLG condition than in the HFLG condition at immediate post-test. One possible explanation is that the association between LFLG and target words was weaker than that between HFLG and target words. Further evidence for this argument is that some lower proficiency participants mismatched a target word with other glosses provided. Another explanation is that the actual degree of accessibility of LFLG was not as high as that of HFLG for lower proficiency participants. Moreover, as is the case with higher proficiency participants, lower proficiency participants made predominant use of the L1 to recall target words (approximately 92%). This is a slightly higher proportion than among the higher proficiency participants, a finding which could be seen as supporting the RHM model (Kroll and Stewart 1994) which assumes that lower proficiency learners rely more on the L2–L1 route than higher proficiency learners.

In contrast to immediate recall, immediate recognition scores from lower proficiency participants showed no significant differences between the HFLG and LFLG conditions. Given that participants' recall but not recognition significantly varied with the frequency of glosses, it would seem that recall is more sensitive to L2 proficiency than recognition.

In both types of L2-gloss conditions, a significant drop in scores was seen between immediate and delayed tests. Such retention loss has repeatedly been reported in previous literature (e.g. Jacobs, Dufon and Hong 1994; Ko 2012; Miyasako 2002; Yoshii 2006). The most likely explanation is that memory degrades due to the 'law of disuse' (Thorndike 1914). Another interpretation is that the declarative memory system, which serves to store lexical knowledge (Ullman 2001, 2005), is vulnerable to forgetting over long retention intervals. In order to ensure that a word can be memorised more durably, as suggested by Jiang (2004), meaning-focused practice and communication are essential in order to integrate explicit lexical knowledge into the linguistic competence system.

Learners' opinions on HFLG and LFLG

As shown in Table 8, most of the participants (72.4%) preferred HFLG to LFLG. This finding is not surprising since learners may be worried about the extent to which lower frequency words can be encoded and comprehended correctly. Another reason is that learners habitually rely on familiar words, and thus they might deem it unnecessary or inefficient to bother with lower frequency glosses. It is worth pointing out, however, that there were 22 participants (13.5%) who did favour LFLGs, indicating that learners' opinions on gloss use were not completely unanimous. It should be noted that correlation analysis showed that L2 proficiency was not significantly correlated with learners' ranked preference for L2-gloss types.

The participants seemed to agree that the main reason for the superiority of HFLG over LFLG was that the former could be encoded and comprehended with ease (Items 1, 2 and 5). We hypothesise a model to illustrate the relationship between glossed words and glosses as shown in Figure 2. A target word is the sign, which needs to be interpreted. An object is another word that grows out of learners' previous experience and knowledge (i.e. glosses). An interpretant mediates the relationship between a sign and an object. According to our questionnaire results, most participants believed that learning new words via HFLGs benefited the object-interpretant process. The assessment results, however,

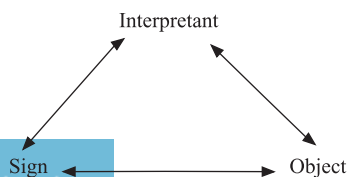

Figure 2. The relationship between glossed words and glosses: a triadic model.

indicate that L2 lexical acquisition among higher proficiency participants was not significantly different between the two L2-gloss conditions. Based on this finding, it seems necessary to reconsider the current pedagogical practice of heavy reliance on high-frequency words.

The present study holds several implications for pedagogical practice. First, language teachers are well advised to insert glosses into texts so as to facilitate learners' incidental lexical acquisition. Long-term retention of lexical items then requires memory consolidation and recycling. Second, LFLGs are not necessarily less effective than HFLGs. Language teachers are best advised to use both types of L2 glosses rather than rely excessively on higher frequency ones. Importantly, teachers should be aware that there is no evidence here to suggest that use of lower frequency glosses necessarily impedes learners' L2 lexical acquisition. More broadly, teacher talk and input materials should be aiming to increase the proportion of lower frequency words used; doing so should assist learners to achieve more ambitious vocabulary targets. Third, L2 proficiency is a factor that deserves consideration in the design and modification of input materials. Ideally, L2 learners should be able to read glossed texts appropriate to their proficiency. Fourth, recall and recognition tests assess different dimensions of a learner's vocabulary knowledge. Teachers should consider which vocabulary test better serves the set pedagogical purpose in assessing individual words.

Conclusion

In this research, we investigated the effects of glossing on L2 lexical acquisition among intermediate L2 learners. Our results were consistent with Jacobs, Dufon and Hong (1994) and Ko (2012): the provision of glosses increases the likelihood of recalling and recognising more words as a by-product of reading, and the gloss advantage was retained in delayed recognition but not in delayed recall. Regarding the effectiveness of HFLG versus LFLG, no significant differences were found, with the exception that lower proficiency learners benefited more from HFLG than LFLG in their immediate recall. The questionnaire results showed that most participants preferred HFLG to LFLG, largely because the former type is less difficult for encoding and comprehension.

The present study has several limitations that we should acknowledge. First, the participants were not allowed to take notes during the learning session, an experimental control that differs markedly from common classroom practice. Second, recycling is important for long-term retention of learned words (Schmitt 2008). In this experiment, the lack of such a consolidation phase again does not reflect typical learning processes. Third, the participants chosen for this study cannot possibly represent the range of L2 learners, and particularly those immersed in the L2 environment, whose incidental vocabulary learning opportunities may well be different. In general, we conclude with two directions for future research: (1) to examine whether and the extent to which the frequency of L2-only explanations affects the outcomes of intentional L2 lexical acquisition and (2) to investigate the extent to which the effectiveness of vocabulary teaching strategies varies with vocabulary test methods.

Notes

1. The lower frequency L2 glosses (LFLGs) had been acquired by the participants.
2. IELTS consists of four sections: listening, writing, reading, and speaking. It takes approximately three hours to complete the whole test. As commented by Stoynoff (2009), IELTS has been empirically demonstrated to have validity, reliability, impact, and practicality. The speaking test was conducted one to one in a quiet room with a qualified and trained examiner. Each IELTS component was scored out of nine bands, and then averaged and rounded to the nearest whole or half band.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Teaching Development Fund of Beijing Jiaotong University, P. R. China.

References

Balota, D.A. and J.I. Chumbley. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance* 10, no. 3: 340–57. DOI: 10.1037/0096-1523.10.3.340.

Barry, C., C.M. Morrison and A.W. Ellis, 1997. Naming the Snodgrass and Vanderwart pictures: effects of age of acquisition, frequency, and name agreement. *The Quarterly Journal of Experimental Psychology Section A* 50, no. 3: 560–85. DOI: 10.1080/783663595.

Cobb, T. 2000. The Compleat Lexical Tutor. http://www.lextutor.ca (accessed 13 June, 2016).

Craik, F.I. and R.S. Lockhart. 1972. Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, no. 6: 671–84. DOI: 10.1016/S0022-5371(72)80001-X.

Day, R.R., C. Omura and M. Hiramatsu. 1992. Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language* 7, no. 2: 541–51.

Field, A. 2013. *Discovering Statistics Using SPSS*. 4th ed. London: Sage.

Forster, K.I. and S.M. Chambers. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12, no. 6: 627–35. DOI: 10.1016/S0022-5371(73)80042-8.

Glanzer, M. and J.K. Adams. 1990. The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16: 5–16. DOI: 10.1037/0278-7393.16.1.5.

Goldinger, S.D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, no. 2: 251–79. DOI: 10.1037/0033-295X.105.2.251.

Hintzman, D.L. 1986. 'Schema abstraction' in a multiple-trace memory model. *Psychological Review* 93, no. 4: 411–28. DOI: 10.1037/0033-295X.93.4.411.

Holley, F.M. and J.K. King. 1971. Vocabulary glosses in foreign language reading materials. *Language Learning* 21, no. 2: 213–19. DOI: 10.1111/j.1467-1770.1971.tb00060.x.

Horst, M., T. Cobb, T. Cobb and P. Meara. 1998. Beyond a Clockwork Orange: acquiring second language vocabulary through reading. *Reading in a Foreign Language* 11, no. 2: 207–23.

Hu, H.-c.M. and H. Nassaji. 2012. Ease of inferencing, learner inferential strategies, and their relationship with the retention of word meanings inferred from context. *Canadian Modern Language Review* 68, no. 1: 54–77. DOI: 10.3138/cmlr.68.1.054.

Huckin, T. and J. Bloch. 1993. Strategies for inferring word meanings: a cognitive model. In *Second Language Reading and Vocabulary Learning*, ed. T. Huckin, M. Haynes and J. Coady, 153–78. Norwood, NJ: Ablex.

Hulstijn, J.H. 1992. Retention of inferred and given word meanings: experiments in incidental learning. In *Vocabulary and Applied Linguistics*, ed. P.J.L. Arnaud and H. Béjoint, 113–25. London: Macmillan.

Hulstijn, J.H., M. Hollander and T. Greidanus. 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, no. 3: 327–39. DOI: 10.1111/j.1540-4781.1996.tb01614.x.

Jacobs, G.M., P. Dufon and F.C. Hong. 1994. L1 and L2 vocabulary glosses in L2 reading passages: their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading* 17, no. 1: 19–28. DOI: 10.1111/j.1467-9817.1994.tb00049.x.

Jiang, N. 2000. Lexical representation and development in a second language. *Applied Linguistics* 21, no. 1: 47–77. DOI: 10.1093/applin/21.1.47.

Jiang, N. 2004. Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal* 88, no. 3: 416–32. DOI: 10.1111/j.0026-7902.2004.00238.x.

Ko, M.H. 2012. Glossing and second language vocabulary learning. *TESOL Quarterly* 46, no. 1: 56–79. DOI: 10.1002/tesq.3.

Kroll, J.F. and A.M.B. De Groot. 1997. Lexical and conceptual memory in the bilingual: mapping form to meaning in two languages. In *Tutorials in Bilingualism: Psycholinguistic Perspectives*, ed. A.M.B.de. Groot and J.F. Kroll, 169–99. Mahwah, NJ: Lawrence Erlbaum Associates.

Kroll, J.F. and E. Stewart. 1994. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33, no. 2: 149–74. DOI: 10.1006/jmla.1994.1008.

Laufer, B. 1989. What percentage of text lexis is essential for comprehension? In *Special Language: From Humans Thinking to Thinking Machines*, ed. C. Lauren and M. Nordman, 316–23. Clevedon: Multilingual Matters.

Laufer, B. 1997. The lexical plight in second language reading: words you don't know, words you think you know, and words you can't guess. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, ed. J. Coady and T. Huckin, 20–34. New York: Cambridge University.

Lv, H., M. Yao and Y. Du. 2005. Effects of word annotation on vocabulary learning in English reading. *Psychological Science* 28, no. 6: 1415–17.
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.
Nassaji, H. 2003. L2 vocabulary learning from context: strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly* 37, no. 4: 645–70. DOI: 10.2307/3588216.
Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
Nation, I.S.P. and D. Belgar. 2007. The vocabulary size test. *The Language Teacher* 31, no. 7: 9–12.
Nöth, W. 2014. The semiotics of learning new words. *Journal of Philosophy of Education* 48, no. 3: 446–56. DOI: 10.1111/1467-9752.12076.
Oldfield, R.C. and A. Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, no. 4: 273–81. DOI: 10.1080/17470216508416445.
Pitts, M., H. White and S. Krashen. 1989. Acquiring second language vocabulary through reading: a replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language* 5, no. 2: 271–75.
Plass, J.L., D.M. Chun, R.E. Mayer and D. Leutner. 2003. Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior* 19, no. 2: 221–43. DOI: 10.1016/S0747-5632(02)00015-8.
Pulido, D. 2009. How involved are American L2 learners of Spanish in lexical input processing tasks during reading? *Studies in Second Language Acquisition* 31, no. 1: 31–58. DOI: 10.1017/S0272263109090020.
Rott, S. 1999. The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition* 21, no. 4: 589–619. DOI: 10.1017/S0272263199004039.
Rott, S., J. Williams and R. Cameron. 2002. The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research* 6, no. 3: 183–222. DOI: 10.1191/1362168802lr108oa.
Rouhi, A. and H. Mohebbi. 2012. The effect of computer assisted L1 and L2 glosses on L2 vocabulary learning. *The Journal of Asia TEFL* 9, no. 2: 1–19.
Schilling, H.E., K. Rayner and J.I. Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition* 26, no. 6: 1270–81. DOI: 10.3758/BF03201199.
Schmidt, R.W. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11, no. 2: 129–58. DOI: 10.1093/applin/11.2.129.
Schmidt, R. 1993. Awareness and second language acquisition. *Annual Review of Applied Linguistics* 13, 206–26. DOI: 10.1017/S0267190500002476.
Schmidt, R. 1994. Deconstructing consciousness in search of useful definitions for applied linguistics. Consciousness in Second Language Learning. *AILA Review* (Special Issue) 11, 11–26.
Schmidt, R. 2001. Attention. In *Cognition and Second Language Instruction*, ed. P. Robinson, 3–33. New York: Cambridge University Press.
Schmitt, N. 2008. Review article: instructed second language vocabulary learning. *Language Teaching Research* 12, no. 3: 329–63. DOI: 10.1177/1362168808089921.
Schmitt, N. and D. Schmitt. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching* 47, no. 4: 484–503. DOI: 10.1017/S0261444812000018.
Stoynoff, S. 2009. Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching* 42, no. 1: 1–40. DOI: 10.1017/S0261444808005399.
Tang, E. and H. Nesi. 2003. Teaching vocabulary in two Chinese classrooms: schoolchildren's exposure to English words in Hong Kong and Guangzhou. *Language Teaching Research* 7, no. 1: 65–97. DOI: 10.1191/1362168803lr113oa.
Thorndike, E.L. 1914. *The Measurement of Ability in Reading, Preliminary Scales and Tests*. New York: Teachers College, Columbia University.
Ullman, M.T. 2001. A neurocognitive perspective on language: the declarative/procedural model. *Nature Reviews Neuroscience* 2, no. 10: 717–26. DOI: 10.1038/35094573.
Ullman, M.T. 2005. A cognitive neuroscience perspective on second language acquisition: the declarative/procedural model. In *Mind and Context in Adult Second Language Acquisition*, ed. C. Sanz, 141–78. Washington, DC: Gergetown University Press.
Watanabe, Y. 1997. Input, intake, and retention: effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition* 19: 287–307. DOI: 10.1017/S027226319700301X.
Yanguas, I. 2009. Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology* 13, no. 2: 48–67.
Yoshii, M. 2006. L1 and L2 glosses: their effects on incidental vocabulary learning. *Language Learning & Technology* 10, no. 3: 85–101.

Appendix. HFLG versus LFLG

	HFLGs			LFLGs		
Target words		GF	GL		GF	GL
Principal	Main	1.00	4.00	Chief	2.00	5.00
Converge	Centre	1.00	6.00	Focus	3.00	5.00
Neglect	To not give enough *care* to something	1.00	4.00	To not give enough *attention* to something	2.00	9.00
Abiding	Something that has *lasted* for a long time	1.00	4.00	Something that has *existed* for a long time	2.00	5.00
Peril	Danger	1.00	6.00	Risk	2.00	4.00
Assent	Agree	1.00	5.00	Permit	3.00	6.00
Prevailing	Usual	1.00	5.00	Common	2.00	6.00
Routine	Normal	1.00	6.00	Regular	2.00	7.00
Unprecedented	That has never *happened* before	1.00	6.00	That has never *occurred* before	2.00	5.00
Elated	Very *happy*	1.00	5.00	Very *joyful*	2.00	6.00
Abatement	Reduction	2.00	9.00	Decrease	3.00	8.00
Swift	Quick	1.00	5.00	Rapid	2.00	5.00

Note: GF: gloss frequency; GL: gloss length.