Bilingualism: Language and Cognition

cambridge.org/bil

The effect of contextual richness on L2 learners' processing and incidental acquisition of non-literal English phrasal verbs: An eye-tracking study

Tianjiao Song[1,2] and Barry Lee Reynolds[2,3]

[1]School of Foreign Languages, Guangdong Ocean University, China; [2]Faculty of Education, University of Macau, Macau SAR and [3]Centre for Cognitive and Brain Sciences, University of Macau, Macau SAR

Research Article

Cite this article: Song, T. and Reynolds, B.L. (2026). The effect of contextual richness on L2 learners' processing and incidental acquisition of non-literal English phrasal verbs: An eye-tracking study. *Bilingualism: Language and Cognition* 1–16. https://doi.org/10.1017/S1366728926101722

Received: 24 September 2025 UTC
Revised: 21 June 2026 UTC
Accepted: 26 June 2026 UTC

Keywords:
non-literal English phrasal verbs (PVs); eye-tracking; contextual richness; L2 incidental vocabulary acquisition; lexical inferencing; vocabulary retention; multi-word units

Corresponding author:
Barry Lee Reynolds;
Email: barryreynolds@um.edu.mo

Abstract

Many studies have used post-tests to examine word-related factors, such as encounter frequency, that influence L2 learners' incidental acquisition of single words. However, less is known about how context affects the processing and acquisition of multi-word units, such as non-literal English phrasal verbs (PVs), which are particularly challenging for L2 learners to master. This eye-tracking study investigated how contextual richness influences L2 learners' processing and acquisition of non-literal English PVs. Fifty-seven Chinese-L1 postgraduate students were randomly assigned to three groups, each reading eight low-frequency PVs in contexts varying in informativeness: *more informative*, *less informative* and *uninformative*. Vocabulary post-tests and delayed post-tests assessed incidental learning gains. Results indicated that *more informative* contexts enhanced form retention, whereas *less informative* contexts facilitated meaning recognition. These findings contribute to cognitive theories of word learning and offer practical insights for L2 vocabulary instruction.

1. Introduction

Incidental vocabulary acquisition occurs when readers pick up new words as by-products during meaning-focused activities, such as reading (Hulstijn, 2003). While being incidental, as with many other learning activities, incidental vocabulary acquisition through reading requires attention and cognitive effort. This is supported by a growing number of eye-tracking studies, which have reported a general positive relationship between the processing time on the new words and incidental learning gains of these words measured by immediate and delayed vocabulary post-tests (Godfroid et al., 2017; Mohamed, 2015; Pellicer-Sánchez, 2016; Pellicer-Sánchez et al., 2020)[1]. However, most of these studies limited their scope to single words with concrete meaning. Less is known about the processing and acquisition of multi-word units that entail idiomatic meaning, particularly non-literal English phrasal verbs (PVs).

Non-literal English PVs, often referred to as verb-particle constructions, are a type of multi-word unit notoriously difficult for L2 learners to acquire (Gardner & Davies, 2007). The predominant learning hurdle rests not only on the complexity of PV forms which are longer than single words but also on the non-literal meaning, the interpretation of which is different from their constituent parts (Garnier & Schmitt, 2016; Siyanova & Schmitt, 2007). For example, the non-literal meaning of *hunker down* (i.e., hold stubbornly to some position) cannot be derived from either *hunker* (sit on one's heels) or *down* (toward a lower position). Compared with single words that entail concrete meaning, non-literal PVs are semantically opaque (Garnier, 2016). Thus, learners need to rely on context to accurately interpret and acquire these non-literal PVs encountered during reading. This leads to the assumption that contextual richness, the informativeness of embedding contexts, should play an important role in how learners process and incidentally acquire non-literal PVs. However, this assumption remains empirically untested. The present study addresses this gap by investigating how contextual richness modulates L2 learners' processing and incidental acquisition of non-literal English PVs.

1.1. Attention, processing and incidental vocabulary acquisition

Incidental vocabulary acquisition can be effectively understood through three interrelated theoretical perspectives: the Noticing Hypothesis, the Type-of-Processing-Resource-Allocation

[1]It should be noted, however, that the strength and consistency of this relationship appear to vary across eye-tracking studies due to differences in reading materials (e.g., short paragraphs in Pellicer-Sánchez, 2016 vs. a 9,000-word novel in Godfroid et al., 2013), target lexical items (e.g., nonwords in Mohamed, 2018 vs. low-frequency real words in Elgort, 2018), and input modality (Wang & Pellicer-Sánchez, 2022). Nonetheless, the present investigation corroborates the predominant finding that longer reading times predict greater incidental vocabulary learning gains.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

(TOPRA) model and the E-Z Reader model. These frameworks emphasize attention as a key driver in the processing and acquisition of words during reading, with the E-Z Reader model (Reichle et al., 2009) specifically highlighting eye movements as measurable indicators of attention distribution while reading.

The Noticing Hypothesis states that attention is a prerequisite for incidental vocabulary acquisition, first by noticing the form of a lexical item (e.g., orthographic properties) and then quickly linking it to meaning to establish a form-meaning connection (Schmidt, 2012). This connection is initially stored in short-term memory and later consolidated into long-term memory for retention (Robinson, 2003). Depth of processing theory (Craik & Lockhart, 1972) further emphasizes that greater initial attention to the form and meaning of a lexical item enhances processing, strengthening its form-meaning connection and supporting learning (Hulstijn & Laufer, 2001). The TOPRA model (Barcroft, 2015) highlighted attention by proposing an interplay between limited cognitive resources and attentional allocation in vocabulary learning. Focusing on meaning enhances meaning acquisition, while attending to orthographic properties supports form acquisition. This model explains the relationship between word processing and acquisition, offering insights into why certain aspects of vocabulary knowledge are acquired more readily than others.

The E-Z Reader model complements these models by providing a cognitive framework to explain how attention is distributed while reading (Reichle et al., 2009). The model posits a direct eye-mind link, with fixation patterns serving as proxies for online processing, and distinguishes between early and late eye-movement measures theorized to index distinct lexical processing stages. This makes it possible to align specific eye-movement measures with the attentional mechanisms postulated by the Noticing Hypothesis and the TOPRA model. Early measures such as First Fixation Duration (FFD) on a word is theorized to index the initial, form-focused encoding (Godfroid et al., 2010). This aligns with the Noticing Hypothesis (Schmidt, 2012), which holds that such initial attention to form is a prerequisite for acquisition to occur. Late measures such as Second-pass Reading Time (SPR) and regressions-in signal strategic processing, including reanalyzing contextual information and integrating inferred meaning (Conklin et al., 2019). These measures indicate meaning-oriented attentional allocation proposed by the TOPRA model (Barcroft, 2015). Total Reading Time (TRT) aggregates all fixations to index overall processing effort, which is found to be associated with the likelihood of acquiring a new word (Elgort et al., 2018). The combined frameworks of the E-Z Reader model, the Noticing Hypothesis and the TOPRA model provide and have been validated by a growing body of eye-tracking studies that consistently reported a general positive relationship between eye-movement measures and incidental vocabulary learning gains (e.g., Godfroid et al., 2017; Mohamed, 2018; Wang & Pellicer-Sánchez, 2022).

Grounded in these integrated theoretical frameworks, the present study employed eye-tracking measures to examine whether contextual richness differentially modulates early form encoding (indexed by early eye-movement measures such as FFD) versus later meaning integration and inferential processes (indexed by late eye-movement measures such as SPR and regressions-in). This approach allowed for a more nuanced understanding of how contextual richness affected the distinct cognitive operations involved in acquiring non-literal PVs.

1.2. Contextual richness in L2 incidental vocabulary acquisition

Few studies using offline measures have examined the effect of contextual richness on incidental vocabulary acquisition through reading. Contextual richness is conceptualized as the amount of information provided in contexts, either in short sentences or stories, for learners to infer the meaning of target words (e.g., Webb, 2008). Focusing on single words, these studies yielded inconsistent findings on whether informative contexts, which provide sufficient information about the meaning of new words, facilitate incidental learning gains (e.g., Webb, 2008; Zahar et al., 2001).

Webb (2008) in a study with 50 Japanese EFL university students found that informative context was conducive to acquiring new word form and meaning. Webb (2008) developed 3 sets of 10 short sentence contexts from graded readers and examined L2 learners' acquisition of 10 single non-words on four measures: meaning recognition, meaning recall, form recognition and form recall. The contextual richness of the short sentences was categorized into two levels: *more informative* and *less informative* based on a four-point contextual rating scale. Webb (2008) found that the participants who encountered the target words in *more informative* contexts demonstrated significantly higher scores on meaning recognition and recall, suggesting that contextual richness may be particularly beneficial for acquiring word meaning.

The facilitative role of contextual richness was also observed in Teng's (2019) study, which examined 180 Chinese EFL undergraduate students' incidental acquisition and retention of single words through reading short sentence contexts. Using Webb's (2008) ratings, Teng (2019) extended Webb's (2008) finding to vocabulary knowledge retention by reporting that the *more informative* contexts facilitated participants' performance on delayed post-tests for all four knowledge types (form recall, form recognition, meaning recall and meaning recognition), particularly for word meaning.

However, there is evidence showing a non-significant effect of contextual richness on L2 learners' incidental acquisition of new words. Zahar et al. (2001) studied 144 Quebec ESL learners reading a short fiction with 30 English words and found no difference in the contextual richness of acquired and non-acquired words. Their study concluded that a *more informative* context did not significantly facilitate incidental vocabulary acquisition. This aligns with Mondria and Boer (1991) and Mondria's (2003) studies, which also found no positive relationship between retention and inferences from *more informative* contexts. Instead, a context with limited information may prompt learners to exert more mental effort in processing new words and verifying inferences through contextual clues via trial and error (Bjork & Kroll, 2015; Mondria & Boer, 1991). This effortful learning may result in learners' greater cognitive engagement with the new words and their meaning, which could benefit the word learning and retention (Mondria & Boer, 1991).

A similar finding was reported by Reynolds (2020), whose hierarchical regression analysis showed that contextual richness was not a statistically significant contributor to incidental vocabulary learning gains. This replication study involved 356 Chinese-L1 undergraduate EFL learners who read Zahar et al.'s (2001) short fiction containing 30 target words, each repeated 2–9 times. Using Zahar et al.'s (2001) test, the study found that participants averagely acquired 9.5% of the target word form through reading. Still, contextual richness appeared to not be a "particularly strong predictor" for incidental learning gains (p.17). Reynolds (2020) suggested that the participants might have paid deliberate attention to target words rather than context, which aided learning. While offline measures were used, the study could not ascertain participants' attention during reading.

More recently, Tadayonifar et al. (2025a) examined the role of contextual support in the intentional learning of figurative English

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

phrasal verbs (PVs). In the study, 85 Iranian EFL learners received definitions for 42 figurative PVs either before or after reading associated texts. Results showed that providing definitions after reading yielded significantly greater learning gains for both PVs' form and meaning. This finding was corroborated in a related study with 98 learners from the same population (Tadayonifar et al., 2025b). These findings, though from an *intentional* learning paradigm, highlight the importance of meaning availability for PV acquisition, providing a conceptual parallel for our study on how contextual richness affects incidental PV learning.

A review of the literature reveals mixed findings on the role of contextual richness in incidental L2 vocabulary acquisition through reading. These inconsistencies likely reflect a methodological limitation, as these previous studies relied on off-line vocabulary post-tests that cannot capture the real-time processing of new words during reading. To address this methodological gap, more recent research used eye-tracking to examine the process of incidental vocabulary learning (e.g., Godfroid et al., 2013; Mohamed, 2015). Primarily based on single words, these studies suggested that contextual richness imposed a small effect on the acquisition of new word forms through reading.

Godfroid et al. (2013) investigated 28 advanced EFL learners' acquisition of 12 non-words through reading short paragraphs with varying contextual richness, operationalized by the presence and position of appositions (non-word + apposition; apposition + non-word). The study found that contextual richness did not significantly impact incidental vocabulary acquisition, as appositions did not affect the probability of recognizing target word forms on a gap-fill post-test. Eye-movement data (e.g., first fixation duration, total reading time) showed participants spent more time on appositions than target words, regardless of whether appositions preceded or followed the target words. The low learning rate (23%) of target words suggests that greater attention to contextual cues (appositions) may not necessarily facilitate the acquisition of new word forms. This finding is supported by Mohamed's (2015) eye-tracking study in which 42 advanced EFL adult learners read a story containing 20 single non-words. The study found that contextual richness significantly predicted the scores of meaning recognition and recall, but not form recognition (Mohamed, 2015). This result paralleled Webb's (2008) findings, suggesting that context had little effect on word form acquisition but significantly influenced word meaning acquisition. Mohamed (2015) attributed the non-significant effect of context on word form acquisition to learners spending less time reading the target words that were more predictable from context.

While these eye-tracking studies suggested that contextual richness benefited meaning acquisition over form, methodological concerns warrant caution. First, neither study controlled for L2 reading ability or vocabulary size, both known predictors of L2 incidental vocabulary acquisition (Pulido, 2003; Tekmen & Daloğlu, 2006) and eye movements during reading (Conklin et al., 2019). Second, the beneficial effect of contextual richness on word *meaning* acquisition observed by Mohamed (2015) was based solely on *single* non-words with concrete meanings (e.g., *window*), which were familiar to advanced L2 learners in his study. The absence of an effect on word form acquisition may be due to the use of completely unfamiliar non-words in both studies. It remains unclear about the role of contextual richness in the processing and incidental acquisition of non-literal phrasal verbs, which involve high-frequency particles known to L2 learners, but convey non-literal meanings as a whole.

1.3. The present study

The present study examined the role of contextual richness on the processing and incidental acquisition of non-literal English PVs through reading. The study was guided by the research question: How does contextual richness moderate the relationship between online processing and the incidental acquisition of non-literal English PVs? This inquiry is grounded in two key premises. First, eye-tracking research consistently shows that longer reading times were associated with greater incidental vocabulary learning gains (e.g., Elgort & Beyersmann, 2024). Second, due to their semantic opacity, non-literal PVs rely heavily on context for interpretation (Garnier, 2016). Together, these premises suggest that contextual richness may affect learners' processing of PVs and, consequently, their incidental acquisition. Accordingly, the following hypotheses were tested:

Hypothesis 1 (Form Recognition): Longer reading times on target PVs will predict higher form recognition scores, with this positive relationship strongest in the *more informative* context, where readily available meaning allows greater cognitive resources to be allocated to encoding PV forms, thereby enhancing form acquisition.

Hypothesis 2 (Meaning Recognition): Longer reading times on contextual clues will predict higher meaning recognition scores, with this effect strongest in the *more informative* context, as richer semantic information enables deeper processing of PV meanings, thereby facilitating meaning recognition.

Hypothesis 3 (Meaning Recall): Longer reading times on contextual clues will predict higher meaning recall scores, with this effect strongest in the *more informative* context, where richer semantic information best supports the stronger form–meaning connections required for meaning retrieval.

2. Methodology

2.1. Participants

The participants were 57 Chinese-L1 postgraduate students (female = 35, male = 22), with an average of 15.45 (SD = 3.15) years of English learning experience.[2] They were upper-intermediate (B2 level) L2 learners with an average self-reported IELTS score[3] of 6.52 (SD = .39, $Min.$ = 5.50, $Max.$ = 7) and an average reading subset score of 6.93 (SD = .40, $Min.$ = 6, $Max.$ = 7). The participants were randomly assigned to one of the context conditions: *more informative* (n = 16), *less informative* (n = 23) and *uninformative* (n = 18). The one-way ANOVA confirmed that the three context groups were statistically similar in L2 reading abilities, $F_{(2,54)} = 2.79$, $p = .70$, L2 vocabulary sizes, $F_{(2, 54)} = 2.26$, $p = .11$, and L2 vocabulary knowledge levels, $F_{(2,54)} = .85$, $p = .44$, as measured by the proficiency tests described below (see Supplementary Appendix 1 for the descriptive statistics).

[2]Of the 81 initial participants, 15 were excluded for not meeting the K2 vocabulary knowledge criterion, and 9 did not complete the experiment, resulting in a final sample of 57 participants included in the analysis.

[3]The participants were first- or second-year postgraduate students in a two-year program, and because IELTS scores are required for admission and remain valid for two years, their self-reported scores were taken as a recent proxy for English proficiency at the time of data collection.

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

2.2. L2 Proficiency measures

Three proficiency tests were administered prior to the reading task: the 14,000 bilingual Mandarin version Vocabulary Size Test (VST) (Nation & Beglar, 2007), the Updated Vocabulary Levels Test (UVLT) (Webb et al., 2017), and the reading subsection of the Cambridge English First for Schools, which assesses English reading proficiency of upper-intermediate learners (B2 Level; IELTS 5–6.5) (https://www.cambridgeenglish.org/exams-and-tests/first/). The VST and the Cambridge reading test measured participants' L2 vocabulary size and L2 reading proficiency, respectively, both known to affect vocabulary learning gains through reading (Pulido, 2003; Tekmen & Daloğlu, 2006) and eye movement patterns (Conklin et al., 2019). The VST scores showed an average accuracy of 46% or lower on the K8 level and above, confirming that participants were unlikely to know the target PVs, which contained low-frequency verbs *beyond* K9. The UVLT scores further showed that the participants had the knowledge of first 2,000 (K2) most frequent English word families based on the test's recommended cutoff criteria (K1: $M = 29.79$, $SD = .41$, $Min. = 29$, $Max. = 30$; K2: $M = 29.39$, $SD = .49$, $Min. = 29$. $Max. = 30$), ensuring that the reading materials (written using K2 vocabulary) would not pose undue lexical difficulty. All three proficiency tests exhibited high reliability (Cronbach's $\alpha > .80$) (DeVellis & Thorpe, 2016).

2.3. Target non-literal phrasal verbs

The targets were eight non-literal English PVs, selected from PVs dictionaries. Great care was taken to control the extraneous variables related to these PVs that are known to affect readers' online reading patterns, including word length, word frequency and frequency of meaning senses (e.g., Kliegl et al., 2004).

2.3.1. Length

The targets were 8-letter PVs (i.e., *dabble in*, *wallow in*), 9-letter PVs (i.e., *squirm out*, *slough off*, *siphon off*, *fizzle out*) and 10-letter PVs (i.e., *hunker down*, *batten down*). All the PVs in each level of context were matched on the length of constituent verbs and particles.

2.3.2. Frequency

First, the constituent verbs in the target PVs are low-frequency words beyond K9 (e.g., *hunker*). Consistent with this, participants' VST scores showed mean accuracy of 46% or lower at the K8 band and above, suggesting limited knowledge of low-frequency vocabulary and making it unlikely that they knew the target verbs. Second, a corpus query demonstrated that the target PVs as a whole (e.g., *hunker down*) are low-frequency and therefore further indicating that they are unlikely to be known by the participants. The frequency count of each target PV adhered to the common query techniques for PV counting corpus analyses, which included the target PVs' different tenses in both contiguous (e.g., *siphons off*, *siphoned off*) and non-contiguous forms (e.g., *siphons it off*, *siphoned it off*) (Gardner & Davies, 2007). A criterion of 10 tokens per trillion in both COCA and BNC was considered as low frequency based on previous studies on compiling PVs lists (Garnier & Schmitt, 2015; Liu, 2011); a *minimum* occurrence of 0.1 tokens per trillion to ensure that the target PVs would provide some benefit for participants' incidental acquisition during reading.

2.3.3. Non-literal meaning

The study focused on the most frequent non-literal meaning of PVs, as it relates to "the most useful semantic usages of the PVs" (Liu & Myers, 2018, p. 33) and represents a particular learning challenge for L2 students (Laufer, 2000). The most frequent non-literal meaning of target PVs was determined by corpus analysis. Using the procedure from Garnier and Schmitt (2015) and Liu and Myers (2018) for creating PVs lists with the most frequent meaning senses, 100 concordance lines were randomly selected from both BNC and COCOA for each target PV, followed by another 100 lines. The non-literal meaning of a target PV in each concordance line was coded independently by two L2 researchers and triangulated with PV dictionaries and WordNet 3.1. Cohen's Kappa values showed high agreement between the raters for all PVs (see Supplementary Appendix 2 for Kappa values). The identified non-literal meaning sense of each target PV was expressed by over 80% of its tokens that denote the non-literal meaning, indicating the high frequency of the non-literal meaning sense (Garnier & Schmitt, 2015).

2.4. Contextual richness

Contextual richness was classified by Webb's (2008) context ratings, which provided a detailed rubric for assessing the informativeness of contexts that enables L2 learners' successful inferencing of unknown word meanings. Drawing on Webb's (2008) context ratings, three levels of context were created: *more informative*, *less informative* and *uninformative*. The meaning of each target PV referred to its most frequent non-literal meaning sense identified in the earlier corpus analysis (see Section 2.3.3.). For the *more informative* context, the target PVs were presented with rich clues drawn from definitional words in dictionary entries (if within K2) or their synonyms (if above K2), allowing readers to reliably infer the intended meaning of the PVs. For the *less informative* context, the target PVs were accompanied by clues consisting of words semantically related to synonyms of the PVs, enabling readers to gain partial knowledge of the PV's meaning. The *uninformative* context provided no information that could help generate any particular meaning of the PVs.

2.5. Sentence contexts creation

Each level of context consisted of three short sentences (Supplementary Appendix 3). To enhance ecological validity and naturalness, these sentences were adapted from authentic corpus examples of each target PV, with the vocabulary restricted to the K2 band to ensure participant comprehension. Minor adjustments were also made to maintain consistency across experimental conditions, as detailed below.

The *first* short sentence and the *third* short sentence were sentence frames, remaining the same for each PV in all three levels of context. These sentence frames ensure the similar positions of the PVs on the screen, away from the beginning or the end of a line, which may cause longer reading time on the target item (Rayner et al., 1995). The *second* sentence involves a target PV, followed by a contextual clue (Table 1). The preceding context before each target PV and sentence frames (i.e., the first and second sentences) were neutral to avoid strong constraints on the target PVs (Schilling et al., 1998). The neutrality of sentence frames, preceding context and classification of context levels were confirmed through two norming studies.

2.5.1. Norming study 1: Contextual neutrality judgments

Contextual neutrality judgments were conducted to ensure that contextual clues were the only source providing information for

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

Table 1. A description of the three levels of contexts with examples of *hunker down*

Target PV and meaning	Three levels of contexts	Example
Hunker down: To hold stubbornly to some position.	*More Informative* Context: Readers are likely to infer the correct meaning of the target PV.	After a careful discussion, the committee finalized the date of the interview with each applicant. Before it started, we told all the candidates to **hunker down** *by not changing their opinions.* The entire interview was open to the public.
	Less Informative Context: Readers are likely to gain partial knowledge of the PV's meaning.	After a careful discussion, the committee finalized the date of the interview with each applicant. Before it started, we told all the candidates to **hunker down** *by not doubting their opinions.* The entire interview was open to the public.
	Uninformative Context: Readers are highly unlikely to infer the correct meaning of the target PV.	After a careful discussion, the committee finalized the date of the interview with each applicant. Before it started, we told all the candidates to **hunker down** *by not planning brief meetings.* The entire interview was open to the public.

Note: In each three-sentence context, the first and the last sentences are sentence frames (underlined). The target PV is bold, and the contextual clues are in italics. The text formatting (bold, italics, underlining) of the contexts in Table 1 is only for reference and was presented in regular typeface in the experimental stimuli.

Please fill in the blank with the word(s) that first came into your mind.

After a careful discussion, the committee finalized the date of the interview with each applicant. Before it started, we told all the candidates to _____. The entire interview was open to the public.

Figure 1. An example of the contextual neutrality judgments.

participants to infer the PVs' meaning. A modified cloze task was used, where the target PVs along with their contextual clues were removed from the three-sentence contexts (see Figure 1 for an example). Sixteen L1 speakers and fifteen L2 learners who were not involved in the main study were asked to provide the missing word(s). The contextual neutrality for a target PV is operationalized as the proportion of participants who successfully predicted the target PV. The results showed that none of the raters provided the target PVs (or their synonyms), suggesting that the sentence contexts were neutral.

2.5.2. *Norming study 2:* Contextual *Richness Rating*

The level of informativeness for each three-sentence context was further evaluated by a norming study, where the same group of raters was presented with three-sentence context for each target PV along with the PV's meaning in the raters' L1 and was asked how likely they could use each context to guess the meaning of the word on a 9-point Likert scale (see Figure 2 for an example). The ANOVA analysis of the ratings revealed that both L1 and L2 raters had significantly different ratings on the three levels of contexts $F_{(2, 42)} = 31.99$, $p < .001$, $\eta_p^2 = .60$, with higher ratings of the *more informative* context (L1: $M = 6.96$, $SD =$. 93; L2: $M = 6.92$, $SD = .71$) than *less informative* (L1: $M = 5.24$, $SD = .98$; L2: $M = 5.82$, $SD = .75$) and *uninformative* context (L1: $M = 4.17$, $SD = .88$; L2: $M = 4.84$, $SD = .90$). Tukey's HSD test for pairwise comparisons showed that ratings were statistically significant (all $ps < .01$), confirming a hierarchical classification of contextual richness.

2.6. Vocabulary measures

Participants' incidental learning gains of the target PVs were measured by three immediate post-tests: form recognition, meaning recall and meaning recognition (Supplementary Appendices 4–7). The multiple-choice form-recognition test was a receptive measure assessing participants' ability to recognize the form of each target PV from its three distractors. The distractors were plausible nonwords that follow common English spelling rules (Nation & Webb, 2011) and were paired with a particle different from the target PV. To prevent participants from confusing the distractors with actual verb + particle combinations, all the distractors were examined by BNC and COCA corpus searches to ensure that the combinations are not used in English.

The meaning recall post-test was a receptive measure of target PV's meaning, requiring participants to provide either the Chinese translation or an English synonym for each PV they had previously selected in the form recognition test. The meaning recognition test evaluated the receptive knowledge of target PVs' meanings by asking the participants to choose the correct meaning from three distractors for each PV.

Vocabulary knowledge retention was evaluated through delayed post-tests, where the target PVs were rearranged to minimize the potential impact of test order on participants' responses. For the multiple-choice form-recognition test and the meaning-recognition test, one point was assigned for each correct response. Answers to the meaning recall test were independently evaluated by two raters proficient in English and Chinese. One point was given for a correct synonym, translation, or definition of each target PV.

2.7. Eye-tracking apparatus

All the reading materials were programed using Experiment Builder and performed through the Eye-Link Portable Duo manufactured by SR Research with a sampling rate of 1,000 Hz. The typeface was in Courier New font, 14 pt., presented on a 21.5-inch computer monitor set at 47 cm (26.8 cm*1.75) from the participants' eyes. The reading materials were presented on a white background to simulate natural reading (Conklin et al., 2019) and were double-spaced to ensure the eye-tracker's accurate recording of the participants' fixations while reading (Godfroid et al., 2013). The participants were seated in front of

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

You will be presented with 8 items. For each item, you will be shown a phrase (always in bold), the meaning of the phrase, and some 3-sentence contexts in which the phrase appears.

Target Phrase: HUNKER DOWN

Meaning: To hold stubbornly to some position 坚持观点

If you came across each of the 3-sentence contexts while reading and did not know the meaning of the phrase, please rate how likely (1 for very unlikely, 9 for very likely) you would be able to use each 3-sentence context to guess the meaning of the phrase.

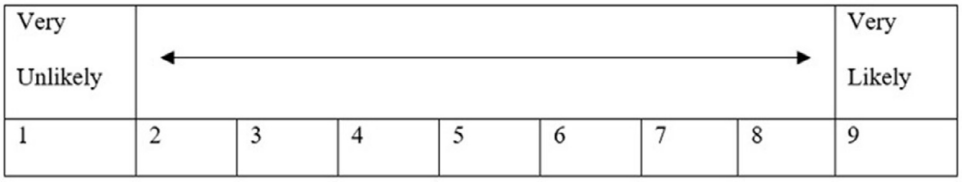


Figure 2. An example of the contextual richness rating.

a computer monitor with their heads placed on a chin rest to optimize accuracy levels and spatial resolution. A 9-point calibration was performed for each participant to ensure the accuracy levels of eye-movement data recording. Validation was performed after each calibration as a further examination on the precision of the eye-tracker.

2.8. Reading task

Participants in each condition viewed 80 experimental screens (8 target PVs × 10 repetitions, with each repetition presenting one 3-sentence context containing a target PV embedded in the context condition to which the participant was assigned, that is, *more informative*, *less informative*, or *uninformative*), 80 unique filler screens (three-sentence contexts with no target PVs, each presented once), 5 practice screens, 18 true/false comprehension questions (each on a separate screen), 3 break pages, one introductory screen and one goodbye page. The 10-time repetition of each target PV was the suggested number of encounters for yielding a "discernable rise" in incidental learning gains through reading (e.g., Pigada & Schmitt, 2006, p. 18). The filler contexts were matched to the experimental contexts in length and structure to increase text variety and prevent participants from detecting patterns in the reading materials, which could otherwise have affected their eye movements. The presentation of all experimental and filler contexts was randomized.

2.9. Eye-movement measures

To examine the effect of contextual richness on the processing and acquisition of target PVs, two separate regions of interest (ROIs) were defined: the PV and the subsequent contextual clue region. Each PV was analyzed as a single ROI because non-literal PVs typically function as holistic syntactic and lexical units, with their contributing verbs and particles combining to convey the overall meaning of the PVs (Quirk et al., 1985). Treating each PV as an individual ROI ensured a comprehensive analysis of the processing of the PVs. Further, defining the PVs and the subsequent contextual clue region as two separate ROIs enabled the analysis of *regressions-in*, a regressive eye-movement moving from the contextual clue region back to the PV, which serves as a direct indicator of learners' use of subsequent context for meaning derivation (Mohamed, 2018).

For the eye-movement measures, first fixation duration (FFD) and first-pass reading time (FPR) were examined as indicators of initial lexical encoding (Clifton et al., 2007; Godfroid et al., 2010) and were found to be associated with word form acquisition (Mohamed, 2018). Second-pass reading time (SPR) and total reading time (TRT) were examined as measures of cognitive processing effort during re-analysis (SPR) and overall cognitive engagement (TRT), respectively (Clifton et al., 2007), both of which have been shown to predict incidental word learning gains (Pellicer-Sánchez, 2016; Yang & Ren, 2025). Regressions-in, defined as returns to the PV after first-pass reading, was examined as an indicator of using subsequent contextual information to verify inferences (Elgort et al., 2018). As all contextual clues followed the PVs, this measure directly reflects how contextual richness influences PV processing.

3. Procedures

The research methods have been reviewed and approved by the ethics review board at the University of Macau (No. MYRG2018-00008-FED). The research process involved three phases: L2

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

proficiency tests, reading task and vocabulary post- and delayed post-tests[4]. In phase 1 (120 min), all the participants read and signed a consent form, completed a demographic questionnaire and sat the L2 proficiency tests in sequential order: the VST, the UVLT and an L2 reading ability test. In phase 2, participants were invited to the eye-tracking lab individually for the reading task and the unannounced vocabulary post-tests. The participants were randomly assigned to one of the three context groups, reading 8 target non-literal PVs embedded in the sentence contexts matching their assigned group. For example, the participants in the *more informative* context group read three-sentence contexts for the 8 target PVs presented in a *more informative* context, along with 80 fillers. The three-sentence contexts were presented in random order for each context group.

To maintain participants' engagement while reading, the participants were informed that some of the screens they read would be followed by reading comprehension questions. A drift correction was performed before the presentation of each screen to ensure accurate eye-movement data collection. A 3-minute break was scheduled after every 108 screens to prevent fatigue from affecting the participants' eye movements while reading. The participants' eye movements on all the screens were recorded by the eye tracker. The total reading session lasted around 60 minutes.

After the participants had completed the reading task, the immediate vocabulary post-tests were conducted in a specific order: form recognition, meaning recall and meaning recognition (Nation & Webb, 2011). Delayed post-tests were conducted 2 weeks after the post-tests to assess vocabulary knowledge retention.

4. Data analysis

4.1. Data cleaning

Prior to analyzing the participants' eye movements, a data-cleaning procedure was conducted. First, a trial-by-trial initial screening found track loss in two filler trials, which did not affect the eye-tracking data analyzed in this study. Second, each fixation on two separate ROIs, PVs and clues, was examined. PVs were considered as an individual ROI based on the rationale that PVs shorter than 80 ms are often conceived as a byproduct of blinks and thus were merged with the neighboring fixations (i.e., 4.3% of the data) (Godfroid, 2020). Fixations longer than 800 ms were excluded as these fixations are often a result of "momentary loss" in attention (i.e., 3.1% of the data) (Conklin et al., 2019, p. 191). The participants' scores on the 18 T/F reading comprehension questions were also analyzed. The results showed that the participants had an average accuracy rate of 98%, M = 17.61, SD = .53, 95% CI [17.47, 17.75]. The one-way ANOVA analysis of the T/F reading comprehension scores showed no significant difference among the three context groups, $F_{(2,54)} = .01$, $p = .10$, $\eta^2 = .00$, suggesting that the three groups comprehended the sentence contexts.

4.2. Data analysis software and packages

All the statistical analyses were conducted with R (version 4.2.1; R Core Team, 2022). This study used a between-subject design where the participants were randomly assigned to one of the three conditions (*more informative* context, *less informative* context, *uninformative* context), reading the same 8 target PVs occurring in each condition, followed by taking vocabulary post- and delayed post-tests to assess their incidental learning gains of these PVs. Mixed-effects modeling was used as it is particularly advantageous for analyzing data with repeated measures (e.g., eye-movement data and vocabulary post- and delayed post-test scores) or hierarchical structures (e.g., PVs were crossed with contexts), accounting for both subject- and item-level variability, reducing residual error and leveraging all data to produce reliable estimates of fixed effects (Westfall et al., 2014). Poisson regression mixed-effects models were fitted using the glmmTMB package in R (Brooks et al., 2017) and was used to analyze count data (regressions-in) (Hutchinson & Holtman, 2005). Logistic mixed-effect models were used to analyze binary data (e.g., form recognition) (Hilbe, 2009). Odds ratios were used to quantify the effect size of individual predictors. The odds ratios larger than 3 or smaller than .33 were considered to be strong (Haddock et al., 1998), equal to 75% probability (Wang & Pellicer-Sánchez, 2022). All the VIF values estimated by "Car package" (Fox, 2018) were smaller than 4, suggesting the absence of collinearity (Larson-Hall, 2010). The outliers of each regression model were identified by a visual check of the residual scatterplot. The data points with absolute standardized residuals of 2.5 standard deviations (i.e., $2.5 < |z|$) above the mean were considered potential outliers and were excluded from the analyses (Godfroid, 2020).

4.3. Model construction

The study hypothesized a moderating role of contextual richness on the processing-acquisition relationship for non-literal PVs during reading. To test this hypothesis, eye-movement measures from two distinct ROIs were analyzed: target PV and clue. For each vocabulary test (e.g., form-recognition post-test), separate logistic mixed-effects models were fitted to assess the predictive role of processing in each ROI. For the PV ROI, the included eye-movement measures were FFD, FPR, SPR and TRT. For the clue ROI, the same four measures (FFD, FPR, SPR, TRT) were analyzed, along with a fifth measure: regressions-in, which specifically captures backward eye movements from the clue region back to the PV region and thus directly indexes the integrative use of contextual information. In each model, contextual richness (context group) and one of the eye-movement measures from the focal ROI were included as fixed effects, with random intercepts for participants and random intercept–slope correlations for contextual richness by item. To test the hypothesized moderation effect separately for each ROI (e.g., PV), an interaction term between context group and the focal eye-movement measure (e.g., Context Group × FFD on PV; Context Group × SPR on PV) was added and was retained only if it significantly improved model fit. The item-related variables (e.g., PV frequency, PV length and clue size[5]) and participant-related variables (e.g., L2 reading ability, the VST and the UVLT) were

[4]We did not administer a pretest primarily because the eye-tracking paradigm is highly sensitive to pre-exposure effects, such as previewing (e.g., Elgort et al., 2024). A pretest occurring immediately before the reading task may direct participants' attention to the target PVs, altering natural fixation patterns and compromising the incidental nature of the learning. It should be noted that recent research has shown that a pretest does not always significantly affect incidental vocabulary learning outcomes when there is a one-week interval between the pretest and the reading task (Puimège et al., 2025). However, such an interval was not present in our design, and therefore the risk of immediate priming remained a relevant concern. While no pretest was included, the participants' unfamiliarity with the target PVs was ensured through standardized vocabulary tests (VST and UVLT) and corpus-based target PVs' selection.

[5]The clue size of each target PV was the same across the three levels of context, whereas the clue sizes were different across the 8 target PVs. Thus, clue size was added to each model as a covariate to improve the model precision by accounting for the potential effect of different clue sizes among the PVs on the outcome variable.

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

added as covariates when they improved model fit. Variables were rescaled where appropriate (Schielzeth, 2010). Model selection followed a backward procedure, prioritizing parsimony, with fit compared via LRT and AIC (Matuschek et al., 2017). For models where a significant interaction was retained, post hoc pairwise comparisons between context groups were conducted, with *p*-values adjusted using Tukey's HSD method to control the family-wise error rate.

5. Results

The descriptive eye-tracking data (Supplementary Appendix 8) indicated that, compared with the other two context groups, the *more informative* context group generally spent more time processing both PVs and clues, reflected in both early and late eye movement measures (e.g., FPR and SPR). The *less informative* context group spent the longest initial processing time (i.e., FFD) on the PVs among the three groups. The *less informative* context group made marginally more regressions-in (from clues to PVs) than the *more informative* context group, though the *uninformative* context group made the most regressions among the three groups. In terms of acquisition outcomes (Supplementary Appendix 9), the *less informative* and *uninformative* context groups had higher scores than the *more informative* context group on the form recognition post-test. However, the *more informative* context group showed better retention of the target PVs' forms (70%), whereas the *less informative* context group demonstrated better retention of the PVs' meanings. Notably, although the *uninformative* context group achieved a high accuracy rate on the form recognition post-test (69%), it consistently underperformed on both immediate and delayed post-tests measuring the PVs' meanings.[6]

5.1. Form recognition

To test Hypothesis 1, which predicted that longer reading times on target PVs would be associated with higher form recognition scores, with this relationship strongest in the *more informative* context, the effects of eye-movement measures on PVs and their interactions with context group (*more informative*, *less informative*, *uninformative*) were examined. Models fitting form recognition post-test scores revealed significant main effects of SPR and TRT. Longer SPR and TRT on target PVs predicted higher odds of form recognition success across all three context groups. No significant interactions emerged between context group and these two eye-movement measures, implying that the facilitative roles of SPR and TRT in recognizing the PVs' form were not dependent on the informativeness of the context in which the PVs were embedded (see Supplementary Appendix 10 for the models).

Further, for the post-test, a significant interaction was found between FFD on PVs and context group, with longer FFD on PVs being associated with the *more informative* context group's lower chance of PVs' form recognition success, $OR = .76$, 95% CI [.65, .88], $p < .001$ (Table 2, Figure 3). In contrast, a positive relationship was found between FFD on PVs and the *less informative* context group's form recognition post-test scores. As FFD on the PVs increased, the *less informative* context group had higher odds form recognition success and significantly outperformed the *more informative* context group, $OR = 1.45$, 95% CI [1.19, 1.76], $p < .001$; $\beta = .37$, $z = 3.76$, $p < .001$). Further, the longer FFD on PVs was associated with the *less informative* context group's higher odds of form recognition success than the *uninformative* context group, though the effect did not reach statistical significance, $OR = .86$, 95% CI [.73, 1.01], $p = .07$.

For the delayed post-test, only a significant main effect of FFD was found, $OR = 1.12$, 95% CI [1.05, 1.20], $p = .001$, suggesting the facilitative effect of FFD on the retention of PVs' form was not modulated by contextual richness (Supplementary Appendix 11). Aside from the eye movement on PVs, eye-tracking data on *clues* were analyzed. The results showed that none of the five eye-movement measures on *clues* significantly predicted form-recognition gains.

5.2. Meaning recognition

To test Hypothesis 2, which predicted that longer reading times on contextual *clues* would be associated with higher meaning recognition scores, with this effect hypothesized to be strongest in the *more informative* context, eye-movement measures on the *clue* and their interactions with context group were examined.

The analyses of meaning recognition post-test scores revealed neither significant main effects of eye-movement measures nor significant interactions with context group (Supplementary Appendices 12–15). Only for the delayed post-test, two eye-movement measures: SPR on clues and regressions-in significantly interacted with contextual richness, respectively (Table 3). The longer SPR on clues was positively associated with the *more informative* context group's meaning recognition scores on the delayed post-test, $OR = 1.36$, 95% CI [1.09, 1.69], $p = .01$. As SPR on clues increased, the *more informative* context group had higher odds of correct meaning recognition on the delayed post-test than *less informative* and *uninformative* context group, respectively (*more informative* vs. *less informative*: $OR = .71$, 95% CI [.54, .94], $p = .02$, $\beta = .34$, $z = 5.60$, $p = .05$; *more informative* vs. *uninformative*: $OR = .90$, 95% CI [.68, 1.18], $p = .44$). No statistically significant difference was found between *less informative* and *uninformative* context group's meaning recognition scores on the delayed post-test when longer SPR was spent on the clues, $OR = 1.26$, 95% CI [.99, 1.60], $p = .06$.

Aside from SPR on clues, a statistically significant interaction was found between context group and regressions-in (Table 4, Figure 4). Regressions-in occurred when participants returned to the target PVs after having read the clues. As regressions-in increased, the *more informative* and *less informative* context groups had significantly higher response accuracy on the meaning recognition delayed post-test than the *uninformative* context group (*more informative* vs. *uninformative*: $OR = .78$, 95% CI [.64, .95], $p = .01$; *less informative* vs. *uninformative*: $OR = .65$, 95% CI [.53, .81], $p < .001$; Tukey post hoc comparisons: *more informative* > *uninformative*: $\beta = .25$, $z = 2.51$,

[6]The *less informative* context group showed marginal improvement in delayed post-test meaning recognition and recall scores. While a testing effect could explain the higher scores, a more likely explanation is that the knowledge acquired from the *less informative* context was unstable, leading participants to correct their responses on the delayed post-tests. This is supported by data showing that 30% of participants in the *less informative* context group corrected their answers on the meaning recognition delayed post-test for the target PV *wallow in*, a higher rate than in the other two context groups. Similarly, 13% of participants in the *less informative* context group corrected their responses to *slough off* on the meaning recall delayed post-test, compared to 6% in the *more informative* and *uninformative* context groups. These patterns suggested that the *less informative* context group's higher delayed post-test scores were likely driven by instability in initial learning, prompting corrections on the delayed post-test.

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

Table 2. Relationship between FFD on target PVs and form recognition post-test scores by context group

Fixed effects	*b*	95% CI	*SE*	*OR*	*z*	*p*
Intercept	.82	[.60, 8.70]	.68	2.28	1.21	.23
Less informative vs. more informative	−.03	[.22, 4.25]	.75	.97	−.04	.97
Uninformative vs. more informative	.20	[.36, 4.18]	.63	1.22	.32	.75
FFD (more informative)	−.28**	[.65, .88]	.08	.76	−3.49	<.001
Less informative × FFD on PV (vs. more informative)	.37**	[1.19,1.76]	.10	1.45	3.76	<.001
Uninformative × FFD on PV (vs. more informative)	.22*	[1.02,1.52]	.10	1.24	2.17	.03
Uninformative × FFD on PV (vs. less informative)	−.15	[.73, 1.01]	.08	.86	−1.81	.07

Random effects		Variance	*SD*
Participant	(Intercept)	2.27	1.50
PV	(Intercept)	2.55	1.60
	Less informative	2.53	1.59
	Uninformative	.92	.96
Marginal R^2 = .01	AIC = 4400.20		
Conditional R^2 = .52	BIC = 4483.20		

Note: **p < .01, *p < .05. More informative = more informative context group, less informative = less informative context group, uninformative = uninformative context group. Baseline: The more informative context group's FFD on target PVs.

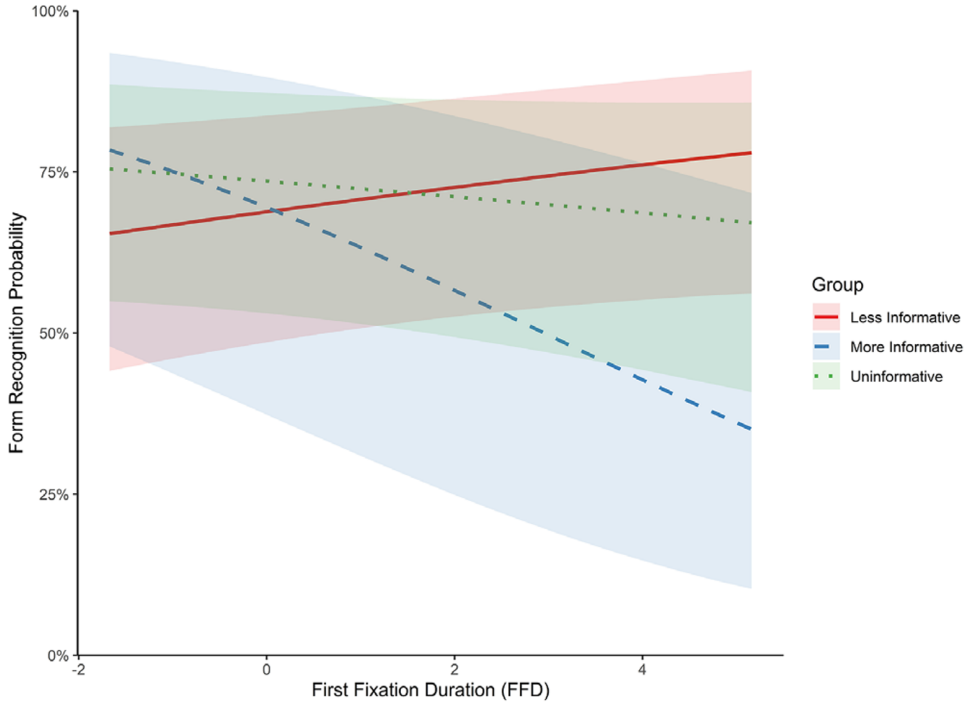


Figure 3. Interaction between FFD on PVs and contextual richness for form recognition post-test scores.

p = .03; *less informative* > *uninformative*: β = .42, z = −3.99, p < .001). Notably, the *less informative* context group had marginally higher meaning recognition delayed post-test scores than the *more informative* context group, *OR* = 1.19, 95% CI [.98, 1.44], p = .08.

Aside from the eye-movement measures on the *clues*, eye-tracking data on *PVs* were analyzed. Results showed that time spent processing target PVs (e.g., SPR, TRT) was associated with lower odds of meaning recognition success on post and delayed post-tests (Supplementary Appendices 16–17). No significant interactions with context group emerged, indicating that the increased attention paid to the PVs *alone* did not facilitate meaning recognition, regardless of contextual richness.

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

Table 3. Mixed-effects models for the relationship between SPR and meaning recognition delayed post-test scores by context group

Fixed effects	*b*	95% CI	*SE*	*OR*	*z*	*p*
Intercept	1.38	[.63, 24.96]	.97	3.97	1.47	.14
Less Informative vs. more informative	.68	[.68, 5.74]	.55	1.97	1.24	.21
Uninformative vs. more informative	−1.04	[.09, 1.45]	.72	.35	−1.44	.15
SPR (more informative)	.31*	[1.09, 1.69]	.11	1.36	2.76	.01
Less informative × SPR on clue (vs. more informative)	−.34*	[.54, .94]	.14	.71	−2.37	.02
Uninformative × SPR on clue (vs. more informative)	−.11	[.68, 1.18]	.14	.90	−.78	.44
Uninformative × SPR on clue (vs. less informative)	.23	[.99, 1.60]	.12	1.26	1.90	.06
Clue size	−3.96*	[.00, .29]	4.33	.02	−2.84	.01
PV length	−1.18*	[.13, .72]	1.40	.31	−2.73	.01

Random effects		Variance	*SD*
Participant	(Intercept)	1.28	1.13
PV	(Intercept)	.56	.75
	More informative	.98	.99
	Less informative	.99	1.00
Marginal R^2 = .18	AIC = 2212.30		
Conditional R^2 = .53	BIC = 2296.90		

Note: *$p < .05$.

Table 4. Relationship between regressions-in and meaning recognition delayed post-test scores by context group

Fixed effects	*b*	95% CI	*SE*	*OR*	*z*	*p*
Intercept	−.74	[.18, 1.28]	.51	.48	−1.47	.14
Less informative vs. more informative	.67	[.70, 5.32]	.52	1.92	1.28	.21
Uninformative vs. more informative	−.76	[.14, 1.52]	.61	.46	−1.24	.21
Regressions-in (more informative)	.19**	[1.07, 1.38]	.07	1.2	2.93	< .01
Less informative × regressions-in (vs. more informative)	.17	[.98, 1.44]	.10	1.19	1.74	.08
Uninformative × regressions-in (vs. more informative)	−.25*	[.64, .95]	.10	.78	−2.51	.01
Uninformative × regressions-in (vs. less informative)	−.42**	[.53, .81]	.11	.65	−3.99	< .001

Random effects		Variance	*SD*
Participant	(Intercept)	1.47	1.21
PV	(Intercept)	1.25	1.12
	Less informative	.81	.90
	Uninformative	1.48	1.21
Marginal R^2 = .06 Conditional R^2 = .51	AIC = 4410.10 BIC = 4493.10		

Note: **$p < .01$, *$p < .05$.

5.3. Meaning recall

Hypothesis 3 predicted that longer reading times on contextual *clues* would be associated with higher meaning recall scores, with this effect strongest in the *more informative* context. To test this hypothesis, eye-movement measures on the *clue* and their interactions with context group were examined.

For the post-test, a significant interaction was found between regressions-in and context group (Table 5, Figure 5). As regressions-in increased, the *more informative* context group demonstrated higher odds of meaning recall success compared to the *less informative* context group, *OR* = .63, 95% CI [.46, .87], β = .46, z = 2.85, p = .004. A similar trend was observed in the *uninformative* context group, although the post hoc analysis showed that the difference was not statistically significant, *OR* = .70, 95% CI [.50, .97], β = .36, z = 2.14, p = .08.

For the delayed post-test, a significant interaction emerged between FFD on clues and context group (Supplementary Appendix 18). Longer FFD on *clues* was associated with higher odds of meaning

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

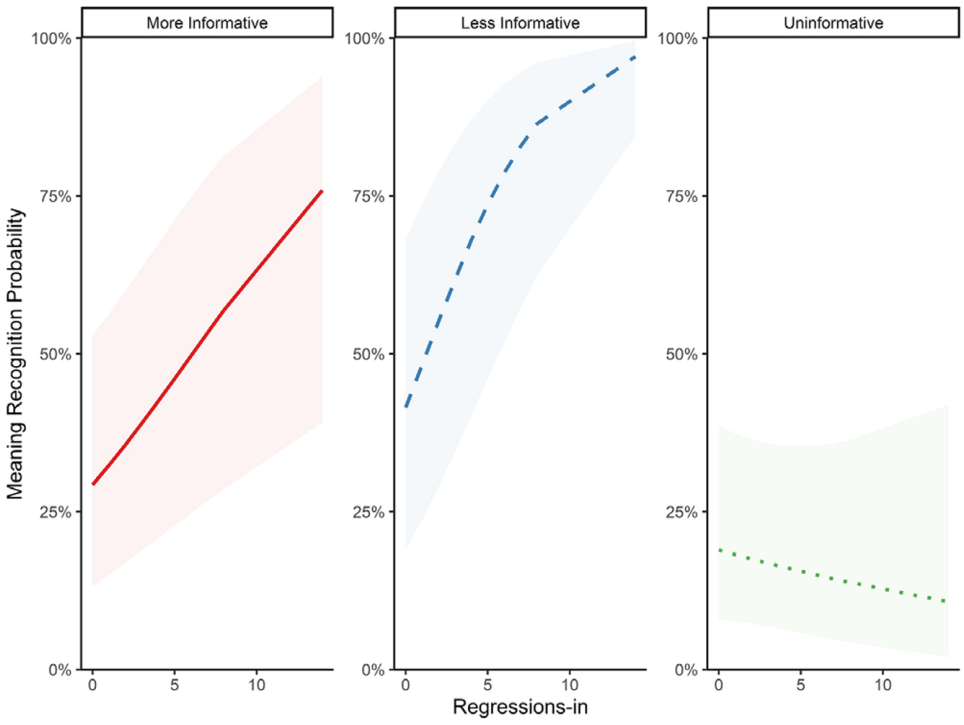


Figure 4. Interaction between regressions-in and contextual richness for meaning recognition delayed post-test scores.

Table 5. Mixed-effects models for the relationship between regressions-in and meaning recall post-test scores by context group

Fixed effects	*b*	95% CI	*SE*	*OR*	*z*	*p*
(Intercept)	−3.56	[.01–.16]	0.87	.03**	−4.12	< .001
Less informative	−1.61	[.01–3.14]	1.41	.20	−1.147	.25
Uninformative	−2.19	[.01–2.17]	1.51	.11	−1.45	.15
Regressions-in	.47	[1.26–2.01]	.12	1.59**	3.93	< .001
Regressions-in × less informative	−.46	[.46–.87]	.16	.63**	−2.85	.004
Regressions-in × uninformative	−.36	[.50–.97]	.17	.70*	−2.14	.032
Random effects				**Variance**		***SD***
Participant		Intercept		7.24		2.69
PV		Intercept		1.78		1.34
		Less informative		8.29		2.88
		Uninformative		9.30		3.05
Marginal R^2 = .05 Conditional R^2 = .79		AIC = 1945.5 BIC = 2028.5				

Note: **p < .01, *p < .05.

recall success in the *more informative* context compared to the *less informative* context group, *OR* = .61, 95% CI [.40, .91], β = .50, z = 5.79, p = .048, and *uninformative* context group, *OR* = .67, 95% CI [.40, 1.13], p = .13. As with meaning recognition, the analysis of eye-movements on the PVs revealed no significant interaction between context group and these eye-movement measures in predicting meaning recall success on the immediate or delayed post-tests.

6. Discussion

While some previous findings have suggested that *more informative* contexts could facilitate the incidental acquisition of word meaning, most of these results were based on learning outcomes of single words measured by vocabulary tests administered at varying points in time after reading (e.g., Teng, 2019; Webb, 2008). Extending the scope to multi-word units, our study used the combined analyses of online (e.g., FFD) and offline data (e.g., vocabulary post-tests) and

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

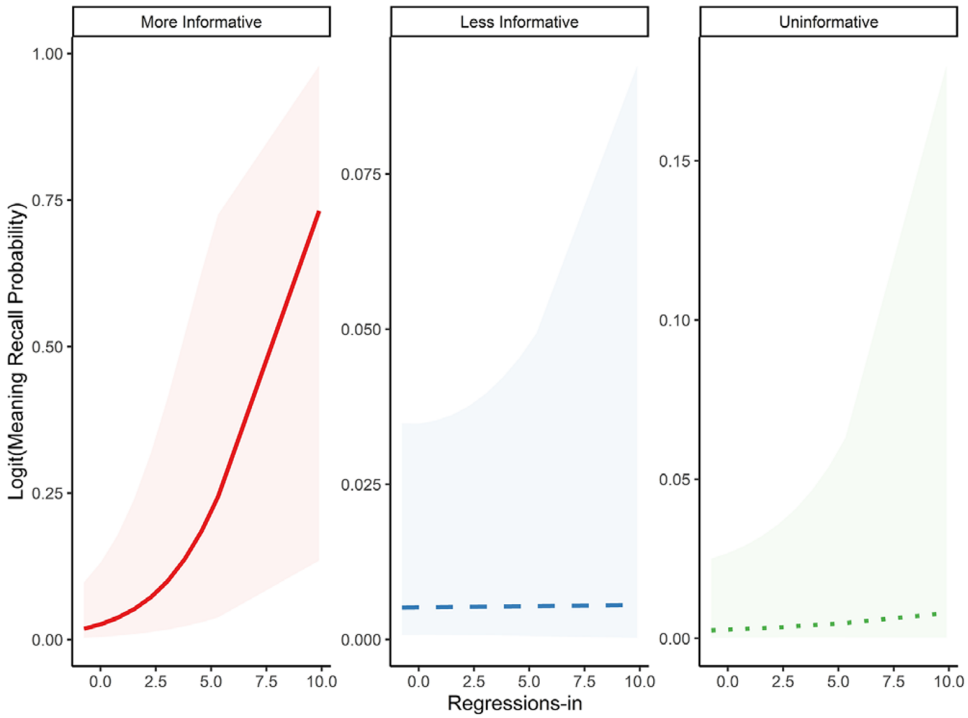


Figure 5. Interaction between regressions-in and contextual richness for meaning recall post-test scores.

found the predictive role of eye movements in the incidental learning gains of non-literal PVs, which generally differed by contextual richness.

6.1. Online processing and form recognition

Hypothesis 1 predicted a positive relationship between reading times on target PVs and form recognition scores, with this effect expected to be strongest in the *more informative* context condition. However, the results revealed a more nuanced pattern: the moderating effect of contextual richness on the processing–acquisition relationship differed between immediate form learning (post-test) and longer-term retention (delayed post-test).

For the post-test, the increased initial attention (FFD) paid to the PVs was more conducive to successful form recognition when the PVs appeared in *less informative* and *uninformative* contexts than in the *more informative* context. This may be due to the inherent limitations of humans' cognitive processing capacity for encoding and retaining input information (Baddeley, 2022; Barcroft, 2015). For the *more informative* context condition, the coexistence of the PVs and rich contextual clues likely increased learners' cognitive load. Learners may have directed more cognitive resources to interpreting contextual clues to infer the idiomatic meaning of the PVs, leaving fewer resources available for noticing and encoding the orthographic forms of these PVs. This pattern aligns with the TOPRA model (Barcroft, 2015), which posits that semantic elaboration may inhibit the encoding of formal word properties when processing demands are high. Conversely, in *less informative* and *uninformative* contexts, the reduced or absent contextual clues may have directed learners' initial attention toward the PVs' forms, increasing the likelihood of noticing and accurately recognizing them on the immediate post-test. This explanation is further illustrated by the descriptive eye-tracking data (Supplementary Appendix 8), which showed that among the three groups, the *more informative* context group spent numerically longer FFD on contextual *clues* than on the *PVs* themselves, suggesting the learners' greater initial attention paid to the PVs' meanings.

Different from the findings of the form recognition post-test, the results of the delayed post-test showed a significant main effect of FFD. Longer FFD on non-literal PVs was found to be associated with L2 learners' better retention of the PVs' form, irrespective of the contextual richness. This finding suggests that the initial lexical processing of non-literal PVs plays a greater role than contextual richness in the retention of the PVs' forms.

FFD, the duration of the first fixation on a region of interest during the first pass, is considered an indicator of noticing, which represents the initial stage of lexical processing (Godfroid et al., 2010). The initial lexical processing is pivotal for memory retention. According to the depth of processing theory (Craik & Lockhart, 1972), the deeper cognitive analysis that occurs during this initial processing helps create robust memory traces, leading to enhanced retention. This helps explain the significantly positive relationship observed between FFD on the target PVs and form recognition delayed post-test scores.

However, the positive effect of FFD on PV form retention was not influenced by contextual richness possibly because initial lexical processing is primarily form-focused. Learners must first recognize a word's form (e.g., orthographic form) before associating it with meaning (Reichle et al., 2009; Schütze, 2017). Focusing on form is crucial for retaining verb-particle constructions, such as PVs, which are longer and syntactically more complex than single words. Therefore, compared to processing contextual clues that were particularly related to the non-literal PVs' meanings, the increased attention to the PVs' forms may have a greater impact on form retention, explaining the significant main effect of FFD on retaining target PV forms.

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

6.2. Online processing and meaning recognition

Hypothesis 2 predicted a positive relationship between reading times on contextual clues and meaning recognition scores, with this effect expected to be strongest in the m*ore informative* context. The results partially supported this hypothesis, though the pattern of support differed across eye-movement measures.

Consistent with the hypothesis, increased SPR on clues in the *more informative* context was associated with better retention of meaning recognition gains. SPR is the sum of all fixations in an ROI when the eyes visit the region for the second time (Conklin & Pellicer-Sanchez, 2016); longer SPR on clues indicated the learners' elaborate processing of contextual clues for extracting the meaning of the target PVs. The advantage of the *more informative* context suggested that the elaborate processing of *more informative* contextual clues could yield greater incidental learning gains in recognizing non-literal PVs' meaning. This finding aligns with the TOPRA model, which associates cognitive processing types with the aspects of word knowledge acquired (Barcroft, 2015). In reading tasks, which are typically meaning-oriented, increased attention to the meanings of new words enhances semantic processing, leading to better retention of word meanings (Barcroft, 2015). Therefore, the longer SPR on *more informative* contextual clues suggests deeper engagement with the meanings of non-literal PVs, improving the retention of meaning recognition gains. By contrast, since the *less informative* context contained partial information about the target PVs' meanings, the elaborate processing of the limited contextual clues did not help learners perceive as much semantic information about the target PVs as processing the *more informative* contextual clues. Likewise, the increased reanalysis of *uninformative* contextual clues did not help learners extract the meaning of the PVs to yield greater meaning recognition gains. This may help explain the superiority of the *more informative* context over the *less informative* and *uninformative* context as SPR on clues increased.

However, a different pattern emerged for regressions-in. Contrary to the hypothesis, increased regressions-in from clues back to the PVs in the *less informative* context led to better meaning recognition retention than in the *more informative* context. This finding reflects the differences in cognitive engagement during regressive eye movements between the two contexts. In the *more informative* context, the meaning of the target PVs was readily accessible, whereas in the *less informative* context, learners had to infer meaning without direct support. Thus, regressions-in within the *less informative* context indicated *more* effortful lexical inferencing, involving active use of contextual information and repeated reanalysis of PV forms to confirm inferred meanings. This is also reflected in the descriptive eye-tracking data, showing a slightly higher frequency of regressions-in made to target PVs from clues in the *less informative* context (M = .91, SD = 1.26) than in the *more informative* context (M = .84, SD = 1.18). This increased engagement with both form and meaning likely strengthened the form-meaning connections of the PVs, thereby facilitating better retention of the meaning recognition gains. This finding corroborates research on lexical inferencing, which suggests that effortful inferencing could strengthen form-meaning connections of new words and enhance long-term retention (Hu & Nassaji, 2012; Rott, 2007). By examining regressive eye movements, our finding empirically supports the role of effortful lexical inferencing, when triggered by *less informative* context, in retaining the meanings of non-literal PVs.

6.3. Online processing and meaning recall

Hypothesis 3 predicted that longer reading times on contextual clues would be associated with higher meaning recall scores, with this effect strongest in the *more informative* context. Consistent with this prediction, more regressions-in in the *more informative* context were found to increase L2 learners' odds of recalling target PV meanings on the post-test compared to the other two contexts. Regressions-in reflect increased processing of informative clues alongside non-literal PV forms, enabling learners to extract sufficient semantic information to establish such robust connections, thereby supporting correct meaning recall on the immediate post-test. This benefit of regressive eye movements for meaning recall is also observed in Prichard and Atkins's (2021) study with Japanese EFL learners, which found a significant correlation between making regressions to review contextual clues and successful meaning recall of new words (r = .30, p = .01). Findings of the present study extend Prichard and Atkins's (2021) observation by suggesting the facilitative role of regressive eye movements in the *more informative* context on facilitating recall of idiomatic meaning of multi-word units.

For the delayed post-test, the FFD on clues significantly predicted L2 learners' retention of meaning recall gains, an effect stronger in the *more informative* context. This finding is in line with depth of processing theory, which posits that the greater initial processing could enhance the mental representation for retaining newly acquired information in long-term memory (Craik & Lockhart, 1972). Thus, the greater noticing of the rich contextual information reflected as longer FFD on the *more informative* contextual clues may yield L2 learners' elaborate processing of the PVs' meaning, which, in turn, enabled the meaning of the target PVs to be stably entrenched in the mental lexicon for successful meaning recall on the delayed post-test.

However, it should be noted that while the FFD effect was statistically significant, the odds ratio was modest (*more informative* vs. *less informative*: OR = 0.61), and overall meaning recall scores on the delayed post-test were relatively low across all three context groups (ranging from 4.13% to 9.25%). This pattern appears to be consistent with previous incidental learning research, which reported an average delayed post-test meaning recall rate of approximately 9% (Webb et al., 2023). The low recall rates can be attributed to the inherent difficulty of the meaning recall task, which required participants to produce the meaning of target PVs rather than simply recognize them. This challenge is amplified for non-literal PV, the idiomatic meaning of which cannot be reliably derived from their constituent parts (Garnier, 2016), and under delayed post-test conditions, where memory consolidation is more vulnerable to attrition. The advantage of the *more informative* context suggests that successful meaning recall of non-literal PVs, even with relatively low gains, depends on the establishment of robust form-meaning connections, and this process demands more semantically enriched encounters.

7. Limitations

This study provides insights into how contextual richness influences L2 learners' processing and incidental acquisition of non-literal PVs through reading, yet several limitations remain. First, our findings are based on L2 learners with the same L1, limiting generalizability. Since L1-L2 similarity influences word acquisition (e.g., Chen & Truscott, 2010), future studies should include learners with diverse L1s. Second, while several learner-related variables were considered, other factors like working memory, which affect word processing and retention (Huang & Jiang, 2022; Yi, 2022), were not examined. Future studies should include working memory to explore its impact on L2 learners' processing and acquisition of non-literal PVs. Third, this study employed a between-subject

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press

design instead of a counterbalanced design because counterbalancing often requires a larger number of items to ensure an adequate number of items for each condition. However, this study only involved a small number of target lexical items (n = 8 PVs), as meticulous criteria were applied to select the target PVs, ensuring minimal influence of word-related properties on participants' eye movements. Implementing a counterbalanced design would have further reduced the number of target PVs per condition. To enhance statistical power, a between-subject design was preferred, allowing each context group to encounter all 8 target PVs. Still, it is important to note that while the between-subject design offered advantages in this study, counterbalancing could better mitigate participant biases or individual differences. Future research is encouraged to employ a counterbalanced design with a larger number of target lexical items, thereby addressing this limitation and further enhancing the validity of the present study's results.

Fourth, considering the attentional sensitivity of the eye-tracking paradigm, no pretest was administered in this study so as not to prime participants' attention toward the target phrasal verbs. Recent evidence indicates that pretesting may not significantly affect offline learning outcomes (Puimège et al., 2025). Future research could examine how the timing of pretests influences both eye-movement behavior and the incidental acquisition of multi-word units. Fifth, to minimize text-related confounds, each group encountered the same PVs in their respective contexts (e.g., *more informative* context). Although filler sentences helped reduce attention bias, repeated contexts might still facilitate learning and affect memory formation of target PVs (Durrant & Norbert, 2010). Future studies should explore the effects of varied and repeated contexts on PVs' acquisition. Sixth, some contextual clues extended across line breaks, as certain PVs required additional words to convey their meanings. Still, each context group encountered the same set of target PVs and their respective clues, ensuring that any influence of contextual clues crossing line breaks was consistent across groups. Future research should examine line breaks' effects on eye movements. Seventh, eye-tracking alone offered limited insight into brain regions involved in vocabulary learning, while neuroimaging could further clarify cognitive processes in incidental acquisition.

8. Conclusion

First, this study reported a positive relationship between processing time and incidental learning gains. These results were in line with previous eye-tracking findings (e.g., Godfroid et al., 2017; Mohamed, 2015; Pellicer-Sánchez, 2016; Pellicer-Sánchez et al., 2020) and provided empirical support for cognitive theories of word learning, such as the depth of processing theory (Craik & Lockhart, 1972), which congruently relate greater cognitive engagement with new lexical items to greater learning. Furthermore, the findings of this study provided empirical support for the Noticing Hypothesis (Schmidt, 2012). Longer FFD was found to be positively associated with the retention of PVs' form, pointing to the importance of noticing in acquiring the forms of multi-word units. Besides, the longer SPR on the *more informative* clues significantly enhanced L2 learners' acquisition of PVs' meaning, lending support to the TOPRA model by suggesting that the L2 learners' increased attention paid to the meaning of a new lexical item during reading is conducive to the acquisition of its meaning.

Second, gauging the effect of contextual richness on the incidental acquisition of non-literal PVs, this study further revealed the complex relationship between lexical inference and retention. One concern with the lexical inference–retention relationship is whether L2 learners still pay enough attention to the form of new words for retention when the meaning of words is sufficiently provided in contexts (Hu & Nassaji, 2012). Examining the L2 learners' eye movements on the non-literal PVs and clues, the study found that making regressions-in in informative contexts (i.e., *more and less informative*) enhanced the L2 learners' meaning recognition gains of non-literal PVs. This finding indicated that in the case of learning non-literal PVs, the L2 learners may still devote a considerable amount of attention to processing both the form and meaning of non-literal PVs, which could enable learners to strengthen the form-meaning connection with the PVs for retention. These findings mirrored the previous observations about the benefit of informative contexts for learning the meaning of new lexical items through reading (e.g., Webb, 2008). More importantly, drawing on the eye-tracking data, this study provided additional explanations for the advantage of informative contexts for acquiring and retaining the meaning of new lexical items through reading.

Third, the findings of this study provided pedagogical implications. Contextual information was found helpful for L2 learners to acquire and retain the form and meaning of non-literal PVs through reading. Further, the reanalysis of the PVs' form and contextual clues, evidenced by regressions-in in *more informative* and *less informative* contexts, could enhance L2 learners' retention of the receptive gains in PVs' meaning (i.e., meaning recognition delayed post-test scores). These findings indicated the benefit of processing both the form and meaning of PVs in acquiring and retaining the meaning of non-literal PVs through reading. Language instructors may consider encouraging L2 learners to draw attention to both non-literal PVs' form and the surrounding contextual information to optimize the likelihood of acquiring the PVs through reading. This could be done by training L2 learners with both form-focused and meaning-focused lexical inferencing strategies (Nation, 2013). Form-focused strategies include encouraging learners to draw attention to the non-literal PVs at the phrase level to increase the chance of noticing the non-literal PVs' form and facilitate learners' correct interpretation of the PVs' meaning (Garnier, 2016). Teachers may also consider using textual enhancement, such as underlining and highlighting the PVs, to enhance the saliency of verb-particle construction. These textual enhancements may increase L2 learners' chance of noticing the non-literal PVs' form as a holistic verb-particle construction and further help learners entrench these constructions in memory (Choi, 2017). Meaning-focused strategies include encouraging L2 learners to be more attentive to text-level information to support their phrase-based analyses of the non-literal PVs, such as actively looking for the words in the contexts that are semantically associated with the meaning of unfamiliar non-literal PVs and verifying their interpretation of the PVs against a wider context to develop the learners' knowledge of PVs through reading.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926101722.

Data availability statement. The data that supports the findings of this study are deposited in OSF at https://osf.io/kuhj7/?view_only=38fb71b9cb9d4c14b6fa7df228ac0b35.

Acknowledgments. The methodology used to collect the data that is reported in the paper was approved by the University of Macau Ethics Review Board under reference No. MYRG2018-00008-FED and was funded by the University of Macau under research grant no. MYRG 2018-00008-FED and MYRG-CRG2024-00034-FED. This article reports findings drawn from part of the first

author's doctoral thesis. The authors sincerely thank Professor Suhad Sonbul, external examiner of the thesis, for her insightful observations and constructive feedback, which contributed to the evolution of the research presented here. We are also grateful to the anonymous reviewers for their careful reading of the manuscript and valuable suggestions. In addition, we wish to acknowledge the scholars whose pioneering work on phrasal verbs and eye-tracking approaches in vocabulary research inspired and informed the study. Finally, we thank the Centre for Cognitive and Brain Sciences for providing the laboratory space that made this research possible.

Competing interests. We have no conflicts of interest to disclose, and all the authors approve and have made contributions to this submission.

References

Baddeley, A. (2022). Working memory and the challenge of language. In J. W. Schwieter & Z. Wen (Eds.), *The Cambridge handbook of working memory and language* (pp. 19–28). Cambridge University Press. https://doi.org/10.1017/9781108955638.004

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. John Benjamins Publishing Company.

Bjork, R. A., & **Kroll, J. F.** (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, **128**(2), 241–252.

Brooks, M. E., **Kristensen, K.**, **Van Benthem, K. J.**, **Magnusson, A.**, **Berg, C. W.**, **Nielsen, A.**, **Hans, J. S.**, **Mächler, M.**, & **Benjamin, M. B.** (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **9**(2), 378–400. https://doi.org/10.32614/RJ-2017-066

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

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

Clifton, C., **Staub, A.**, & **Rayner, K.** (2007). Eye movements in reading words and sentences. In *Eye movements* (pp. 341–371). Elsevier.

Conklin, K., & **Pellicer-Sanchez, A.** (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, **32**(3), 453–467. https://doi.org/10.1177/0267658316637401

Conklin, K., **Pellicer-Sánchez, A.**, & **Carrol, G.** (2019). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Craik, F. I. M., & **Lockhart, R. S.** (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning Verbal Behavior*, **11**(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

DeVellis, R. F., & **Thorpe, C. T.** (2016). *Scale development: Theory and applications* (4th ed.). Sage Publications.

Durrant, P., & **Norbert, S.** (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, **26**(2), 163–188. https://doi.org/10.1177/0267658309349431

Elgort, I., & **Beyersmann, E.** (2024). Do reading times predict word learning? An eye-tracking study with novel words. *Studies in Second Language Acquisition*, **46**(4), 1282–1297. https://doi.org/10.1017/s0272263124000585

Elgort, I., **Brysbaert, M.**, **Stevens, M.**, & **Van Assche, E.** (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, **40**(2), 341–366. https://doi.org/10.1017/S0272263117000109

Elgort, I., **Wetering, R. v.**, **Arrow, T.**, & **Beyersmann, E.** (2024). Previewing novel words before reading affects their processing during reading: An eye-movement study with first and second language readers. *Language Learning*, **74**(1), 78–110. https://doi.org/10.1111/lang.12579

Fox, J., & **Weisberg, S.** (2018). *An R companion to applied regression*. Sage Publications.

Gardner, D., & **Davies, M.** (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, **41**(2), 339–359. https://doi.org/10.1002/j.1545-7249.2007.tb00062.x

Garnier, M. (2016). *English phrasal verbs: Usage, knowledge, acquisition* (Publication Number 10595024) [Ph.D., The University of Nottingham (United Kingdom)]. ProQuest Dissertations & Theses Global. England. https://libezproxy.um.edu.mo/login?url=https://www.proquest.com/dissertations-theses/english-phrasal-verbs-usage-knowledge-acquisition/docview/1896365788/se-2?accountid=12206

Garnier, M., & **Schmitt, N.** (2015). The PHaVE list: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, **19**(6), 645–666. https://doi.org/10.1177/1362168814559798

Garnier, M., & **Schmitt, N.** (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, **59**, 29–44. https://doi.org/10.1016/j.system.2016.04.004

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge, Taylor & Francis Group.

Godfroid, A., **Ahn, J.**, **Choi, I.**, **Ballard, L.**, **Cui, Y.**, **Johnston, S.**, **Lee, S.**, **Sarkar, A.**, & **Yoon, H. J.** (2017). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism*, **21**(3), 563–584. https://doi.org/10.1017/S1366728917000219

Godfroid, A., **Boers, F.**, & **Housen, A.** (2013). An eye for WORDS. *Studies in Second Language Acquisition*, **35**(3), 483–517. https://doi.org/10.1017/S0272263113000119

Godfroid, A., **Housen, A.**, & **Boers, F.** (2010). A procedure for testing the noticing hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Cognitive processing in second language acquisition* (pp. 169–197). John Benjamins Publishing Company.

Haddock, C. K., **Rindskopf, D.**, & **Shadish, W.** (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, **3**(3), 339–353. https://doi.org/10.1037/1082-989X.3.3.339

Hilbe, J. (2009). *Logistic regression models*. Chapman & Hall/CRC.

Hu, H. C. M., & **Nassaji, H.** (2012). Ease of inferencing, learner inferential strategies, and their relationship with the retention of word meanings inferred from context. *The Canadian Modern Language Review*, **68**(1), 54–77. https://doi.org/10.3138/cmlr.68.1.054

Huang, L., & **Jiang, J.** (2022). The role of working memory in unfamiliar word processing across proficiency levels: Evidence from eye movements. *Journal of Cognitive Psychology*, **34**(5), 607–621. https://doi.org/10.1080/20445911.2022.2088763

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). Wiley.

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

Hutchinson, M. K., & **Holtman, M. C.** (2005). Analysis of count data using poisson regression. *Research in Nursing & Health*, **28**(5), 408–418. https://doi.org/10.1002/nur.20093

Kliegl, R., **Grabner, E.**, **Rolfs, M.**, & **Engbert, R.** (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, **16**(1–2), 262–284. https://doi.org/10.1080/09541440340000213

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica*, **54**(2), 186–196. https://doi.org/10.1111/1467-9582.00059

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, **45**(4), 661–688. https://doi.org/10.5054/tq.2011.247707

Liu, D., & **Myers, D.** (2018). The most-common phrasal verbs with their key meanings for spoken and academic written English: A corpus analysis. *Language Teaching Research*, **24**(3), 403–424. https://doi.org/10.1177/1362168818798384

Matuschek, H., **Kliegl, R.**, **Vasishth, S.**, **Baayen, H.**, & **Bates, D.** (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, **94**, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mohamed, A. A. (2015). *The roles of context and repetition in incidental vocabulary acquisition from L2 reading: An eye movement study* [Doctoral thesis, Michigan State University].

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

Mondria, J. A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: An experimental comparison of the "meaning-inferred method" and the "meaning-given method". *Studies in Second Language Acquisition*, **25**(4), 473–499. https://doi.org/10.1017/S0272263103000202

Mondria, J.-A., & **Boer, M.** (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, **12**(3), 249–267. https://doi.org/10.1093/applin/12.3.249

Nation, I. S. P., & **Beglar, D.** (2007). A vocabulary size test. *The Language Teacher*, **31**(7), 9–13.

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

Nation, I. S. P., & **Webb, S.** (2011). *Researching and analyzing vocabulary*. Heinle.

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., **Conklin, K.**, & **Vilkaitė-Lozdienė, L.** (2020). The effect of pre-reading instruction on vocabulary learning: An investigation of L1 and L2 readers' eye movements. *Language Learning*, **71**(1), 162–203. https://doi.org/10.1111/lang.12430

Pigada, M., & **Schmitt, N.** (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, **18**(1), 1–28. https://doi.org/10.64152/10125/66611

Prichard, C., & **Atkins, A.** (2021). The effect of different vocabulary coping strategies on incidental vocabulary acquisition: An eye tracking study. *The Reading Matrix: An International Online Journal*, **21**(2), 1–16. https://readingmatrix.com/files/25-s394q48o.pdf

Puimège, E., **Caltabellotta, E.**, & **Peters, E.** (2025). Pretesting effects on incidental L2 vocabulary learning through reading. *Studies in Second Language Acquisition*, **47**(3), 774–793. https://doi.org/10.1017/S0272263125100910

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, **53**(2), 233–284. https://doi.org/10.1111/1467-9922.00217

R Core Team. (2022). R: A language and environment for statistical computing (Version 4.2.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Quirk, R., **Greenbaum, S.**, **Leech, G.**, & **Svartvik, J.** (1985). *A comprehensive gammar of the English language*. Longman.

Rayner, K., **Raney, G. E.**, & **Pollatsek, A.** (1995). Eye movements and discourse processing. In R. F. Lorch & J. E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 9–35). Lawrence Erlbaum Associates, Inc.

Reichle, E. D., **Warren, T.**, & **McConnell, K.** (2009). Using E-Z reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, **16**(1), 1–21. https://doi.org/10.3758/PBR.16.1.1

Reynolds, B. L. (2020). The effects of nonce words, frequency, contextual richness, and L2 vocabulary knowledge on the incidental acquisition of vocabulary through reading: More than a replication of Zahar et al. (2001) & Tekmen and Daloğlu (2006). *IRAL-International Review of Applied Linguistics in Language Teaching*, **58**(1), 75–102. https://doi.org/10.1515/iral-2015-0115

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631–678). Wiley.

Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, **57**(2), 165–199. https://doi.org/10.1111/j.1467-9922.2007.00406.x

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, **1**(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Schilling, H. E. H., **Rayner, K.**, & **Chumbley, J. I.** (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, **26**(6), 1270–1281. https://doi.org/10.3758/BF03201199

Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. In W. M. Chan,. S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (Eds.), *Perspectives on individual characteristics and foreign language education* (Vol. **6**, pp. 721–737). De Gruyter.

Schütze, U. (2017). *Language learning and the brain: Lexical processing in second language acquisition*. Cambridge University Press.

Siyanova, A., & **Schmitt, N.** (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL - International Review of Applied Linguistics in Language Teaching*, **45**(2), 119–139.

Tadayonifar, M., **Elgort, I.**, & **Siyanova-Chanturia, A.** (2025a). Contextual learning and retention of phrasal verbs: The effects of definition placement and typographic enhancement. *Studies in Second Language Acquisition*, **47**(1), 157–180. https://doi.org/10.1017/S0272263124000718

Tadayonifar, M., **Siyanova-Chanturia, A.**, & **Elgort, I.** (2025b). The effects of definition placement and lag of retrieval practice on contextual learning and retention of phrasal verbs. *Language Learning*, **75**(4), 1153–1184. https://doi.org/10.1111/lang.12706

Tekmen, E. A. F., & **Daloğlu, A.** (2006). An investigation of incidental vocabulary acquisition in relation to learner proficiency level and word frequency. *Foreign Language Annals*, **39**(2), 220–243. https://doi.org/10.1111/j.1944-9720.2006.tb02263.x

Teng, F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *Language Learning Journal*, **47**(2), 145–158. https://doi.org/10.1080/09571736.2016.1244217

Wang, A., & **Pellicer-Sánchez, A.** (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*, **72**(3), 765–805. https://doi.org/10.1111/lang.12495

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, **20**(2), 232–245. https://doi.org/10.64152/10125/66826

Webb, S., **Sasao, Y.**, & **Ballance, O.** (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *International Journal of Applied Linguistics*, **168**(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Webb, S., **Uchihara, T.**, & **Yanagisawa, A.** (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, **56**(2), 161–180. https://doi.org/10.1017/S0261444822000507

Westfall, J., **Kenny, D. A.**, & **Judd, C. M.** (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, **143**(5), 2020–2045. https://doi.org/10.1037/xge0000014

Yang, Y., & **Ren, H.** (2025). Attention and outcomes across learning conditions in L2 vocabulary acquisition: Evidence from eye-tracking. *Journal of Eye Movement Research*, **18**(3), 21. https://doi.org/10.3390/jemr18030021

Yi, W. (2022). Processing of novel L2 compounds across repeated exposures during reading: A growth curve analysis. *Applied PsychoLinguistics*, **43**(3), 551–579. https://doi.org/10.1017/S0142716422000017

Zahar, R., **Cobb, T.**, & **Spada, N.** (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, **57**(4), 541–572. https://doi.org/10.3138/cmlr.57.4.541

https://doi.org/10.1017/S1366728926101722 Published online by Cambridge University Press