Studies in Second Language Acquisition **42** (2020), 7–32
doi:10.1017/S0272263119000561

Research Article

CONTEXTUAL WORD LEARNING IN THE FIRST AND SECOND LANGUAGE

DEFINITION PLACEMENT AND INFERENCE ERROR EFFECTS ON DECLARATIVE AND NONDECLARATIVE KNOWLEDGE

Irina Elgort*

Victoria University of Wellington

Natalia Beliaeva

Victoria University of Wellington

Frank Boers

Western University

Abstract

Access to definitions facilitates the learning of word meanings when novel words are encountered in reading. However, the memorial costs and benefits of inferring word meanings from context, compared to seeing definitions of unfamiliar words before reading, are not yet well understood. We conducted two experiments with adult L1 (English) and L2 (Chinese) readers to investigate whether the development of declarative and nondeclarative word knowledge benefits more when definitions are supplied before reading (errorless treatment) or after reading (trial-and-error treatment). Study participants encountered 90 target vocabulary items three times in short informative texts under errorless or trial-and-error conditions and entered their meaning inferences immediately after reading each text. Posttreatment, we evaluated participants' declarative knowledge of the target items using a meaning generation (recall) task and nondeclarative knowledge using a self-paced reading task. The trial-and-error treatment followed by definitions resulted in a superior declarative and nondeclarative knowledge, compared to the errorless treatment, for L1 and L2 readers. Inference errors affected the development of declarative but not nondeclarative knowledge, and the trajectory of the development of

This research was supported by a research grants from Victoria University of Wellington. We gratefully acknowledge the assistance of Murielle Demecheleer and Mark Toomer in the data collection. We thank Susan Gass and Luke Plonsky for their editorial support, and three anonymous SSLA reviewers for their helpful comments on earlier versions of this paper.

*Correspondence concerning this article should be addressed to Irina Elgort, Victoria University of Wellington, P.O. Box 600, Wellington, New Zealand. E-mail: Irina.Elgort@vuw.ac.nz

Copyright © Cambridge University Press 2019

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

nondeclarative knowledge was different for L1 and L2 readers. We interpret these findings in terms of the declarative and nondeclarative memory processes underpinning contextual word learning.

INTRODUCTION

Is it more beneficial to view definitions of unfamiliar words before or after encountering them in reading? When a definition of a new word is presented before it is encountered in reading, the likelihood of a contextual inference error is greatly reduced; we will therefore refer to this as an errorless treatment. When an unknown word is first encountered during reading, meaning inference errors may occur; we will refer to this as a trial-and-error treatment. A case can be made for both approaches. Research with native speakers (particularly those with memory impairments) shows that avoiding errors at the encoding stage of learning new words may be beneficial (Baddeley & Wilson, 1994). However, benefits have been reported when word meanings are initially inferred from context and then verified by definitions. In the present study, we investigate the effect of *errorless* and *trial-and-error* approaches in contextual word learning (CWL). Following Nation (2001), we adopt a more comprehensive conceptualization of what it means to know a word; going beyond declarative knowledge of form-meaning mapping (e.g., ability to explain the meaning of a given word form), we also attempt to measure nondeclarative knowledge of use (e.g., ability to fluently access and integrate the meaning of a word into an ongoing meaning construction during reading).

LEARNING NEW WORDS FROM READING

New words can be learned during reading, both in the native (L1) and additional (L2) language (Elgort, Brysbaert, Stevens, & Van Assche, 2018; Elgort & Warren, 2014; Godfroid et al., 2018; Joseph, Wonnacott, Forbes, & Nation, 2014; Pellicer-Sánchez, 2015; Pellicer-Sánchez and Schmitt, 2010; Saragi, Nation, & Meister, 1978). A number of studies suggest that, although readers become familiar with the form of a word relatively quickly, the development and integration of semantic knowledge takes longer and, in some cases, remains incomplete after multiple contextual encounters (Elgort et al., 2018; Elgort & Warren, 2014; Waring & Takaki, 2003; Webb, 2007).

A key process in CWL is inferring word meanings from context. A meaning of a novel word (once noticed in the written input) may be covertly inferred by the reader in the course of making sense of the text or it may be overtly inferred through shifting attention to the novel word form and deriving meaning from its immediate (clause/sentence) and global context (Perfetti & Stafura, 2014; Rieder, 2002; Van Dijk & Kinstch, 1983). Word knowledge established using these two types of inferences may be dissociable. Deliberate, overt inferences are more likely to affect the quality of declarative (explicit) word knowledge, that is, the kind of knowledge measured by readers' ability to generate an accurate definition for a given word. Covert inferences implicated in implicit learning are likely to contribute to the development of nondeclarative[1] (e.g., procedural and

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

implicit) word knowledge, that is, the kind of knowledge commonly measured by online processing times (Paradis, 1994; Reber, 2008, 2013).

What mechanisms underpin the development of the knowledge of word meanings in contextual word learning? An instance-based memory model (Bolger, Balass, Landen, & Perfetti, 2008; Reichle & Perfetti, 2003) proposes that each contextual encounter with a word establishes an episodic memory trace of that word plus its context. Multiple encounters with the word in diverse contexts reinforce overlapping memory traces resonating with previous episodes, eventually leading to an abstraction of a core meaning independent of individual learning contexts (i.e., establishing a semantic representation). Exposure to definitions in conjunction with CWL is particularly beneficial in the development of semantic knowledge (e.g., Fraser, 1999; Stahl & Fairbanks, 1986). This is because definitions represent a word's core semantic features that may resonate with episodic memory traces from individual contextual encounters, facilitating the emergence of the core meaning (Bolger et al., 2008). Whether definitions are presented before or after contextual encounters with a word, the establishment of its meaning representation is likely to be accelerated compared to CWL without definitions (Fraser, 1999; Jenkins, Stein, & Wysocki, 1984; Stahl & Fairbanks, 1986). However, to the best of our knowledge, no studies have systematically compared CWL from pre- and postexposure to definitions, combined with forced meaning inference generation.

ERRORLESS AND TRIAL-AND-ERROR WORD LEARNING

Meaning inferences for unfamiliar words during reading in L1 and L2 may be incorrect (Bensoussan & Laufer, 1984; Frantzen, 2003; Kelly, 1990; Laufer, 1997; Mondria, 2003). Erroneous inferences generate false memory candidates that create a need for memory-taxing postretrieval procedures to verify the appropriateness of the retrieved information in a given context (Rodríguez-Fornells, Kofidis, & Muente, 2004). Although an assumption that this may interfere with learning seems logical, the effect of erroneous inferences on CWL is not particularly well understood. There is evidence that uncorrected incorrect meaning inferences during reading negatively affect the development of declarative word knowledge (Carpenter, Sachs, Martin, Schmidt, & Looft, 2012; Hulstijn, 1992) but there are very few studies that have investigated whether explicit erroneous inferences affect the development of nondeclarative knowledge (Elgort, 2017; Rodríguez-Fornells et al., 2004) and, to the best of our knowledge, no CWL studies systematically manipulated errorless and trial-and-error learning treatments with and without feedback.

Research into the effect of error avoidance (also referred to as *errorless learning*) is mostly conducted under paired-associate learning conditions. In these studies, errorless learning tends to produce superior results on word recall measures (e.g., object naming, stem completion), particularly with severely memory impaired adult participants (Baddeley & Wilson, 1994) and children (Warmington, Hitch, & Gathercole, 2013). Conversely, trial-and-error training with healthy participants yields better results in testing studies when learning involves information retrieval (see Clare & Jones, 2008; Middleton & Schwartz, 2012 for critical reviews); a positive relationship has also been observed between retrieval difficulty and learning (e.g., Karpicke & Roediger, 2007, 2008). Potts and Shanks (2014) reported positive effects of trial-and-error learning on

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

meaning recognition of rare English words and words in an unfamiliar language, Euskara. They conjectured that active inference generation may facilitate the encoding of corrective feedback.[2] Superior learning for the trial-and-error over errorless method was also reported by Cyr and Anderson (2015) when conceptual processing was encouraged (e.g., participants guessed *rose* or *orchid* before being presented with the correct answer, *tulip*), but not when lexical processing was prioritized in a stem completion task (e.g., tu_____). The authors argued that conceptual guesses may have acted as an elaboration on semantic relation (Craik & Tulving, 1975) providing a memory cue for the target word (Knight, Ball, Brewer, DeWitt, & Marsh, 2012). The extent to which conceptual errors benefit memory may be related to the extent to which the errors overlap semantically with correct answers (Cyr & Anderson, 2018; Huelser & Metcalfe, 2012).

A neurophysiological study by Rodríguez-Fornells et al. (2004) showed that trial-and-error learning is associated with deeper memory processes at encoding and postretrieval processes of contextual appropriateness verification, which are likely to benefit the outcomes of CWL (see also Potts, Davies, & Shanks, 2018). Rodríguez-Fornells et al. observed an increased late positive component (LPC) for correctly guessed words in trial-and-error compared to errorless learning, suggesting that items in the trial-and-error condition had been encoded more deeply. In addition, a greater error-related negativity (ERN) observed for correct and plausible incorrect responses in the trial-and-error condition compared with correct responses in the errorless condition evidenced increased postretrieval error checking in trial-and-error learning.

The studies reviewed so far compared errorless and trial-and-error learning under paired-associate learning conditions and may not be optimally informative for an investigation of word learning in reading that involves contextual inference making and synthesis of multiple diverse cues. When new words are encountered in reading, erroneous meaning inferences are more likely to be conceptual than lexical errors because word meanings need to be integrated into context, to make sense of the text (but see Laufer, 1997, for a discussion of deceptive transparency and L2 reading comprehension). In highly semantically constraining contexts, even incorrect meaning inferences are likely to overlap semantically to some degree with the correct meanings.

ERRONEOUS INFERENCES AND CWL

A key insight from CWL studies that examined the effect of meaning inferences and feedback on word learning is that corrective feedback mediates the effect of incorrect inferences. In an CWL study by Carpenter et al. (2012) university students of German learned L2 words in one of the four learning conditions: inferring + English feedback, inferring + German feedback, inferring + no feedback, or marginal gloss. Their knowledge of the target words was tested using a translation task immediately after reading and four days later. While the gloss group scored the highest on the immediate posttest, on the delayed test, the inference group with English feedback received the highest scores. Importantly, inference errors were seldom repeated by participants who received feedback; this suggests that inference errors made during reading did not impair students' ability to encode correct meanings of the target words when feedback was provided.

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

A finding that accuracy feedback can improve CWL outcomes, especially when there is an increased risk of errors, was also reported by Frishkoff, Collins-Thompson, Hodges, and Crossley (2016). In this L1 study, adult participants were instructed to generate one-word definitions for rare English words presented in high- or low-constraint sentence contexts. Active meaning generation required participants to attend to the target word in each trial, thus increasing task engagement and necessitating retrieval, but it also increased the risk of errors in learning (Metcalfe & Kornell, 2007). To counteract negative effects of errors, participants in the feedback group received accuracy feedback (i.e., *correct*, *partially correct*, or *incorrect*) after each meaning generation attempt; participants in the control group did not receive feedback. The feedback group showed greater meaning generation accuracy on the last learning trial and on the synonym judgment task when target words were presented in a mix of high- and low-constraint sentences that increased uncertainly.

Initial evidence regarding the effect of inference accuracy on the development of declarative and nondeclarative word knowledge in reading was reported by Elgort (2017). In this study, Chinese learners of English encountered novel L2 words three times in supportive sentence contexts. They generated a meaning inference for each target item and then reviewed their definitions. Incorrect meaning inferences had a negative effect on the development of declarative word knowledge (operationalized as a meaning generation task) but not on the development of nondeclarative (implicit) knowledge measured in a mixed-modality masked repetition priming task. This finding suggests a possible dissociation between the effect of incorrect meaning inferences on the development of these two kinds of word knowledge from reading. This conclusion is aligned with memory models (declarative/procedural; explicit/implicit) in which redundant analogous knowledge may be acquired in parallel in the two systems, but the learning trajectories and nature of the knowledge learned in the two systems is often quite different (Paradis, 1994; Reber, 2008, 2013; Ullman & Lovelett, 2018).

Elgort's study, however, has a number of limitations. Firstly, participants learned target words in sentence (rather than continuous text) contexts, which could have made learning easier (Laufer & Shmueli, 1997). The use of longer contexts may offer further insights into the effect of inference errors on the development of declarative and nondeclarative knowledge. Secondly, Elgort (2017) investigated the effect of errors in L2, but not L1 reading. Finally, the researcher did not deliberately manipulate errorless versus trial-and-error treatments; item definitions always followed contextual meaning inferences, that is, the effect of errors was investigated only in the trial-and-error mode followed by feedback. Our study addresses these limitations by presenting target lexical items in short connected texts and providing definitions either before or after the CWL treatment.

THE PRESENT STUDY

In the present study, L1 and L2 participants read short texts that contained three occurrences of target vocabulary items (pseudowords) under three conditions. In the errorless condition,[3] participants were presented with short dictionary-type definitions of the target items before reading the texts. In the trial-and-error condition, participants first read the text containing the target item and entered a meaning inference prior to seeing its definition (this condition afforded error correction through the provision of feedback). In

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

the control condition, participants read the text with the target items and entered a meaning inference but no definition was provided. Because robust vocabulary knowledge is underpinned by both declarative and nondeclarative memory systems (Dóczi & Kormos, 2016; Hulstijn, 2002), we used two types of measures: word reading times (RTs) in an online self-paced reading task (as a measure of nondeclarative knowledge) and accuracy of responses in an offline meaning generation task (as a measure of declarative knowledge).

RESEARCH QUESTIONS

The present study investigates the following research questions:

1. Is errorless or trial-and-error treatment with definitions more beneficial in CWL?
   1.1. Is errorless or trial-and-error treatment more beneficial for declarative knowledge in CWL?
   1.2. Is errorless or trial-and-error treatment more beneficial for nondeclarative knowledge in CWL?
2. Does inference correctness affect CWL?
   2.1. Do incorrect inferences affect both declarative and nondeclarative knowledge?
   2.2. Do incorrect inferences affect contextual word knowledge differently under errorless and trial-and-error learning conditions?
3. Does the learning method (trial-and-error vs. errorless) affect L1 and L2 CWL in a similar or different way?

We ran two experiments to investigate these questions: Experiment 1 was conducted with native speakers of English and Experiment 2 with native speakers of Chinese.

EXPERIMENT 1

METHOD

Participants

Study participants were 55 undergraduate university students[4] (38 females), who self-identified as native speakers of English. Their average age was 21 (SD = 3.35). They completed an online English LexTALE test (Lemhöfer & Broersma, 2012) prior to being accepted into the study (a score of 75% or higher was required to be accepted). Their average LexTALE score was 94% (SD = 6.42%). Thirty-six were studying subjects in the humanities and law; the rest were enrolled in science, engineering, or commerce degrees. They received a $50 voucher for their participation.

MATERIALS

Target Vocabulary Items and Their Definitions

In the learning treatment, participants encountered 90 vocabulary items (*targets*) in reading passages; 87 of the items were pronounceable pseudowords and 3 were very low-frequency real words. The pseudowords were used to ensure no prior knowledge of targets without the need for pretests that could affect initial word processing in reading.

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

The targets were six or seven letters long, with either none or one orthographic neighbor, selected from MCWord database (Medler & Binder, 2005). The items included 66 nouns, 16 verbs, and 8 adjectives (Appendix A). The meanings of the items were either in the domain of building/construction (n = 45) or health/medicine (n = 45). These two topics (variable name = *Theme*) are specialized and incorporate many low-frequency words unfamiliar not only to L2 but also to L1 speakers. The theme of building/construction was expected to be somewhat less familiar to university students than the theme of health/medicine (this was later confirmed in informal postexperimental debriefs and reflected in the accuracy of responses in the meaning inference task in the learning treatment). The meanings assigned to the pseudowords were those of the real words they replaced in the reading passages used in the treatment phase; their definitions were slightly modifying definitions of the corresponding real words. The definitions were, on average, 14 words long (SD = 4.6); 98.5% of the words in the definitions were within the first 6,000 word families[5] of English (source: word frequency lists based on the British National Corpus [BNC] and Corpus of Contemporary American English [COCA], online: www.victoria.ac.nz/lals/about/staff/paul-nation#vocab-lists).

Reading Passages

Each target occurred three times in a reading passage.[6] The passages were, on average, 120 words long (SD = 10). They were slightly modified original texts sourced from COCA and the Internet, with minor edits made to ensure that correct meaning inferences were possible. To facilitate comprehension, 96.5% of the words in the reading passages were within the first 6,000 most-frequent word-families of English. Targets appeared in brackets in the passages, emulating a scenario in which a new word is noticed.

Learning Treatments

During the learning phase (Figure 1), the following three treatments (variable name = *Treatment*) were counterbalanced in a within-participant experimental design: errorless (ERL; target definition shown before reading the passage and making a meaning inference); trial-and-error (TAE; target definition shown after reading the passage and entering a meaning inference); and control (CTRL; reading the passage and entering a meaning inference, target definitions not shown). Although we refer to responses in the ERL treatment as meaning inferences, they represent a combination of the initial memory encoded from an item's definition and its additional semantic features derived from context. All targets appeared in all three treatment conditions across three experimental lists, and each participant was assigned to the next available list on their arrival. Passage presentation order was randomized by the E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) for each participant.

Posttests

Participants' nondeclarative knowledge was estimated using an online self-paced reading (SPR) task (Bordag, Kirschenbaum, Tschirner, & Opitz, 2015; Just, Carpenter, & Woolley, 1982; Kim, Crossley, & Skalicky, 2018; Marsden, Thompson, & Plonsky,

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

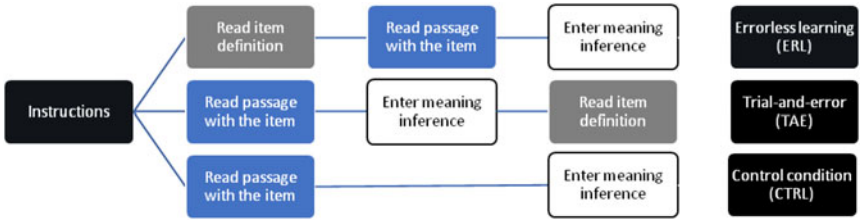


FIGURE 1. Schematic representation of the three learning treatments.

2018). This is not to suggest no involvement of declarative knowledge in reading (both memory systems are often involved in language processing and use, see Ullman & Lovelett, 2018); instead, following Just et al. (1982), we argue that SPR, comparable with the gaze duration measure in eye-movement studies, necessitates the use of nondeclarative knowledge.

In SPR, participants read texts one word at a time, controlling word display duration. Time needed to read each word is recorded. SPR is used to investigate "incremental integration of words and phrases into developing sentence representations" (Witzel, Witzel, & Forster, 2012, p. 106). RTs recorded on predefined areas of interest are used to evaluate ease/difficulty associated with the processing of semantic, syntactic, or pragmatic violations or ambiguity (Marsden et al., 2018). For example, RT on a word following a novel word tends to be longer than that following a familiar word (Just et al., 1982). Because SPR involves online processing of words in context, does not require explicit decision making, and reduces the number of available reading strategies, it is used as a measure of nondeclarative (procedural, implicit) knowledge (Marsden et al., 2018).

A noncumulative, moving-window SPR task (Just et al., 1982) was coded and presented using E-Prime. Each trial contained a short informative text comprising two sentences; the first sentence contained a word or phrase related in meaning to the target presented in the second sentence, for example, "Poppy closed the door of the old piece of furniture. This hacord used to contain her mother's clothes." The meaning of the target, *hacord*, was "a piece of furniture used for storage." The targets appeared at least two words away from the end of the sentence and never at the start. The target and the word following it ($n + 1$) were main interest areas (Bordag et al., 2015; Marsden et al., 2018). Longer RTs on $n + 1$ (often referred to as a *spillover* region) can indicate difficulties with higher-level processes, such as inference making or integrating the target's meaning into the construct at the phrase, sentence, or text levels (Just et al., 1982). In eye-movement studies of reading, regressive eye movements are associated with encountering a semantic or syntactic inconsistency or a need to correct an inappropriate interpretation. Because regressive eye movements are not possible in SPR, inconsistencies and errors are likely to cause longer RTs on the spillover region instead.

Declarative knowledge of meaning was measured using a meaning generation (MG) task, in which participants typed meaning explanations of the targets. Although both recall and recognition tests are used to measure declarative knowledge (Haist, Shimamura, & Squire, 1992), previous errorless and trial-and-error learning studies found

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

that recall measures (such as MG) are better able to differentiate between the effect of these two encoding approaches (Warmington & Hitch, 2014). Also, MG tests readers' ability to access a word's meaning when its written form is presented, which is the kind of knowledge needed in reading. The MG posttest was coded and delivered using *Qualtrics* (www.qualtrics.com). Each target was presented in a short nondefining sentence (e.g., He saw the *adimper*), one at a time, in a random order generated by the software.

Procedure

The experiment was administered over three sessions. Participants completed the procedure individually in a soundproof lab. The first two sessions were conducted on two consecutive days or, at most, two days apart. Sessions one and two were structured in the following way: the learning treatment (Figure 1) was followed by a break and an intervening task (a counting backward task and a language background survey on Day 1; a visual working memory symmetry task on Day 2); followed by the SPR and MG (immediate) posttests. Sessions one and two took 1.5–2 hours each. The building/construction targets (n = 45) were presented on Day 1 and the health/medical targets (n = 45) on Day 2. The third, delayed, session was conducted 2 weeks later; it took no longer than 60 minutes; participants completed SPR and MG posttests with all 90 targets.

Treatment Procedure

Participants were seated in front of the computer and instructed to read each passage through once, with understanding. In the TAE treatment, participants read the text first and then were prompted to enter a meaning inference for the target, after which a new screen with the target definition was shown. In the ERL treatment, participants saw the target definition screen first, which was replaced by the reading text screen, after which participants were prompted to enter a meaning inference for the target. In the CTRL treatment, participants read the text first and then were prompted to enter a meaning inference for the target. Participants were instructed to base their inferences on the text in which targets occurred. Targets were presented in brackets in the passage; the passage remained onscreen while participants typed meaning inferences into a text field below the passage. Their responses and time-on-task were recorded by the software. The following presentation deadlines were used: 100 seconds for reading passages, 30 seconds for definitions, and 30 seconds for the participants to enter a meaning inference. There were two short breaks in the procedure, one after each 15 passages. In a practice set completed before the main procedure, participants read six passages (on an unrelated topic), in which mid- or low-frequency words were used as targets, instead of pseudowords. An experimenter observed participants' behavior and provided feedback when needed.

Posttest Procedure

The immediate SPR posttest was presented to the participants as another reading task (rather than a test) after the completion of the initial learning treatment of the 45 targets on Day 1 and Day 2. They were instructed to read as naturally as possible, at a speed they would normally read. On the initial screen, all words were masked, with dashes in place

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

of letters within words and words separated by spaces. Participants pressed the *Right Arrow* key on the keyboard to reveal each consecutive word. The word was revealed immediately after a keypress, while the preceding word was masked again. RTs were recorded from the word onset until the keypress. After each text, participants were presented with a comprehension statement; they pressed the *Y* key if it was correct and the *N* key if it was incorrect. The statements did not contain the targets. The purpose of the comprehension check was to encourage participants to read with understanding. The procedure started with a practice set of six texts on an unrelated topic containing the items from the practice set of the learning treatment.

In the MG procedure, participants typed an explanation or synonym of the target and indicated whether they had seen or heard it prior to the study; there was no time limit per item. If they were unable to recall the meaning of an item, they could leave the field blank and move to the next item. The MG posttest was always administered after SPR. This is because, to the participants, SPR was just another reading task while MG was a posttest involving an explicit attempt at meaning retrieval, which could have changed the way targets were processed in the SPR. Although encountering targets in SPR prior to the MG test constitutes an additional contextual learning episode (be it without explicit meaning inference or feedback), this learning was the same for all three treatments (ERL, TAE, and CTRL). Thus, an additional contextual exposure to targets in SPR should not have changed the relative effect of *Treatment* on the participants' declarative knowledge.

In all delayed posttests, the targets (n = 90) were grouped into blocks by theme, and the order of the blocks was rotated for each next participant. The delayed posttests were completed by 53 participants.

Data Analysis

Statistical analyses were conducted to examine the effect of *Treatment* and inference accuracy (*Guess.ACC*) on the development of declarative and nondeclarative knowledge of the targets. The data analysis was conducted using the *lme4* package (version 1.1-17; Bates, Maechler, Bolker, & Walker, 2015) in R. Participants and items were entered in the models as crossed random effects. A minimally adequate statistical model was fitted to the data, using a stepwise variable selection and the likelihood ratio test for model comparisons (Baayen, Davidson, & Bates, 2008). The final models contained random slopes supported by the data (i.e., parsimonious mixed models based on Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The inclusion of random slopes in the model allows to better control for the Type I error rate. To further control for the Type I error rate, the function *glht* in the R package *multcomp* (Hothorn, Bretz, & Westfall, 2008) was used to obtain multiplicity-adjusted *p*-values and the corresponding confidence intervals (CI). On completion of the statistical analyses, the fidelity of the results reported in the paper was further checked by performing an adjustment for false discovery (for the MG and SPR analyses separately), which did not change any of our main findings.

Accuracy of initial inferences during the learning treatment and accuracy of responses in the meaning generation posttest was scored as either correct (1) or incorrect (0). Logistic mixed models were fitted to the binary data. We calculated odds ratios (OR) and standardized effect sizes *d* (Chinn, 2000) for the fixed effects in all analyses (Appendix B and C). We are following Brysbaert and Stevens (2018) in interpreting effect sizes in the

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

analysis of the declarative knowledge tests, with a typical effect size in similar psychology studies between $d = .4$ and $d = .3$. The modeling of the MG scores was conducted on the combined immediate and delayed posttest data; a variable, *Session* (with two levels: *immediate* and *delayed*) was added to the analysis. This is because the declarative knowledge posttests were likely to be perceived as the same test in both sessions, even though some knowledge attrition was expected in the delayed test.

In the RT analysis of the SPR data, linear mixed-effects models were fitted to inverse-transformed response times (−1000/RT) on the target and $n + 1$. A minimum *a priori* outlier removal was performed (i.e., only extreme outliers were removed); the final regression models were subjected to model criticism; potentially harmful outliers (i.e., data points with standardized residuals exceeding 2.5 standard deviations) were removed and the model was refitted (Baayen, 2008; Baayen et al., 2008; Brysbaert & Stevens, 2018). Effect sizes in the RT analyses were calculated following recommendations made by Brysbaert and Stevens (2018) for linear mixed-effects analyses (based on Westfall, Judd, & Kenny, 2014). Note that effect sizes in mixed-effects analyses tend to be smaller compared to effect sizes in the analysis of variance (ANOVA); that is, standardized effect sizes are commonly around $d = .1$ in experiment-based RT studies.[7] We also note that typical effect sizes reported in a meta-analysis of SPR studies of grammar with L1 and L2 participants (Avery & Marsden, 2019) are quite small, even for within-participant designs (e.g., $d = .20$ for sensitivity to morphosyntactic information during L2 SPR and $d = .28$ during L1 SPR). The results of the immediate and delayed SPR were analyzed separately; this is because the first SPR was presented to the study participants as a continuation of the reading procedure, whereas the second (delayed) SPR was a posttest. Therefore, participants could adopt different approaches to completing the SPR task in the immediate and delayed sessions.

RESULTS

Participants' responses were scored by two researchers, using a scoring protocol that was initially applied to a random sample of meaning inferences and MG responses, and adjusted accordingly. Correct answers contained core meaning features specific to the target; responses that were too generic (e.g., *a lofty tool*, *a medical institution*) were scored as incorrect. An initial agreement of 91% for the inference accuracy and 98% for the MG accuracy. The final scores were assigned after discussing differences and agreeing on the ratings.

Meaning Inferences during the Learning Treatment

Across all treatments, the mean inference accuracy was 88%. Responses shorter than one second were excluded from the analysis (13 data points; 0.26%), because such short response times do not represent a valid attempt at explaining (typing) the meaning. The mean inference time was 15.62 seconds (SD = 7.83 seconds). Inference accuracy was significantly better under the errorless learning (ERL) treatment than under the control (CTRL) and trial-and-error (TAE) treatments ($z = 12.51$ adjusted $p < .001$; $z = 12.65$ adjusted $p < .001$, respectively; Appendix B, Tables B.1a and B.1b). Across the three treatments, meaning inferences in the health set were more accurate than those in the

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

building set by 3% ($z = 2.26$ $p < .05$). Slower inference generation times were observed on more accurate responses ($z = 3.34$ $p < .001$).

Meaning Generation

There was a statistically significant two-way interaction ($z = -3.00$ $p < .01$) in the analysis of MG scores between Treatment and Inference accuracy (Appendix B, Tables B.2a and B.2b). For correct meaning inferences (Figure 2), the MG accuracy for TAE was significantly higher than for CTRL ($z = 3.25$ adjusted $p < .01$) and ERL ($z = 3.41$ adjusted $p < .01$). For incorrect meaning inferences, the MG accuracy for TAE was significantly higher compared with CTRL ($z = 4.74$ adjusted $p < .001$) but not ERL ($z = 1.58$ adjusted $p = .43$), although it was still higher numerically. The MG accuracy for ERL was not significantly better than for CTRL regardless of whether inferences were correct. Incorrect inferences were most detrimental to learning when definitions were not available.

There were also main effects of Treatment, Inference accuracy, and Session (immediate/delayed) in this analysis. The odds of an accurate MG response to TAE items were 3 times the odds of that for CTRL items (95% CIs: 1.88,4.77; $d = .61$, i.e., greater than .4), but the accuracy of responses to ERL items was not significantly different from CTRL and the effect size of this comparison was relatively small ($d = .27$). Inference accuracy was a significant predictor of the MG accuracy ($z = 4.78$ $p < .001$ $d = .63$), with about 10% MG accuracy improvement when initial inferences were correct. The

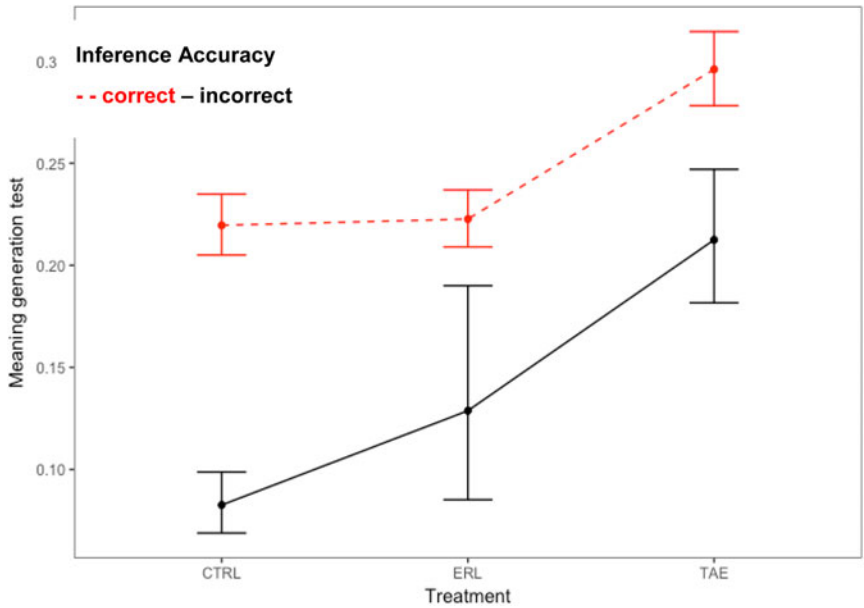


FIGURE 2. The estimated interaction between *Treatment* and *Inference accuracy* (Guess.ACC) in the analysis of L1 MG scores (fit and 95% pointwise confidence intervals). *Note*: CTRL = CWL without definitions, ERL = errorless treatment; TAE = trial-and-error treatment followed by definitions.

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

participants' MG response accuracy was about 14% higher on the immediate than the delayed posttest ($z = 8.63$ $p < .001$ $d = .45$).

In an additional analysis, we compared MG scores for items correctly inferred in the ERL treatment (representing "true" errorless learning) with items correctly inferred in the TAE treatment (TAE.ACC.1) and items incorrectly inferred in the TAE treatment (TAE.ACC.0). The MG accuracy for correctly inferred ERL items and incorrectly inferred TAE items was not different ($z = 0.16$, adjusted $p = .99$), but meaning generation was significantly more accurate for correctly inferred TAE items than correctly inferred ERL items ($z = 3.92$ adjusted $p < .001$) (Figure 3).

Self-Paced Reading

The response accuracy to the comprehension questions in the SPR task was 93.52%; it was 93% on the immediate SPR and 94% on the delayed SPR. This confirms that the SPR texts were well understood by the L1 readers.

Immediate SPR

After trimming extreme outliers (13 data points, 0.26%), the mean RT on the target was 332 ms (SD = 213) and it was 362 ms (SD = 179) on the $n + 1$. There was no effect of Treatment or Inference accuracy in the analysis of RTs on the target. RTs were significantly faster (by about 31 ms; $t = -4.65$ $p < .001$ $d = 0.30$) on the health than building items.

In the $n + 1$ analysis, RTs were predicted by Treatment, Theme, and RTs on the preceding target but not by Inference accuracy (Appendix B, Table B.3a and B.3b). RTs on $n + 1$ were faster for TAE than for CTRL ($t = -2.74$ adjusted $p = .017$) but the predicted difference was very small (7.2 ms) and so was the effect size ($d = .07$, below $d = .1$) even for a mixed-effects analysis of RTs. RTs on $n + 1$ for TAE were about

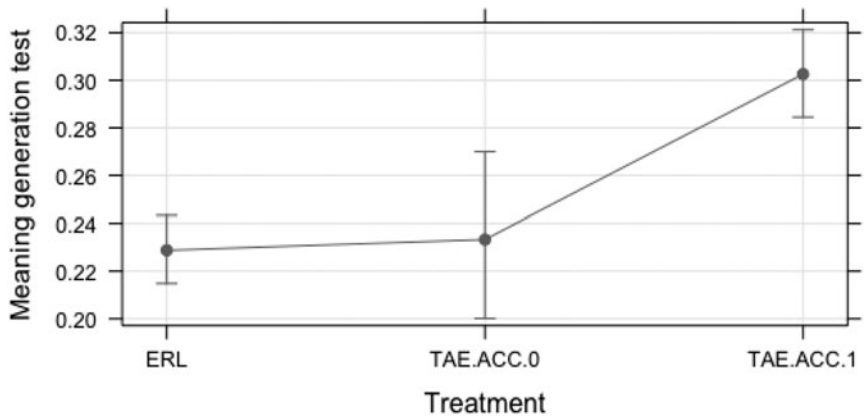


FIGURE 3. The estimated results for the correctly inferred items encountered in the errorless treatment (ERL), correctly inferred trial-and-error items (TAE.ACC.1), and incorrectly inferred trial-and-error items (TAE.ACC.0) in the analysis of L1 MG scores (fit and 95% pointwise confidence intervals).

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

5.4 ms faster than those for ERL; this contrast did not reach statistical significance ($t = -2.07$ adjusted $p = .095$). There was no significant difference between the $n + 1$ RTs for the ERL and CTRL treatments. RTs on $n + 1$ were faster for the health than building theme ($t = -5.03$ $p < .001$) and when RTs on the preceding target were faster ($t = -5.60$ $p < .001$).

We also compared RTs for the correctly inferred (encoded) ERL items with RTs for correctly and incorrectly inferred TAE items, on the target and $n + 1$. RTs on the target did not differ between the treatments. RTs on $n + 1$ were significantly faster for correct TAE items than for correct ERL items ($t = -2.57$ adjusted $p = .025$ $d = .07$); there was no difference between incorrect TAE items and correct ERL items.

Delayed SPR

There was no reliable effect of Treatment or Inference accuracy in the delayed SPR analysis. The mean RTs in the delayed SPR were faster on both the target (270 ms) and $n + 1$ (287 ms) compared to the immediate SPR. Because reading comprehension accuracy did not decrease in the delayed compared to immediate SPR, we conjecture that the overall processing of contextually learned items improved over time, regardless of the treatment. In an additional analysis comparing RTs on correctly inferred ERL items and all TAE items, there was no effect of Treatment on the target but an interesting pattern was observed on $n + 1$: $n + 1$ RTs for incorrectly inferred TAE items were faster than those for correctly inferred (encoded) ERL items ($t = -2.14$ adjusted $p = .077$ $d = .08$). This suggests that contextually learned items that had been incorrectly inferred in the trial-and-error treatment followed by definitions, over time, may have established better semantic representations than items correctly inferred in the errorless treatment.

DISCUSSION

The results of Experiment 1 are straightforward. Firstly, better knowledge of meaning was established in trial-and-error learning followed by definitions than in errorless learning when a definition was shown before encountering the word in reading. This was the case regardless of whether explicit meaning inferences in the errorless learning treatment were correct or incorrect. Errorless learning offered no advantage over trial-and-error learning (with or without definitions) on measures of both declarative and nondeclarative knowledge. Better declarative knowledge of meaning was observed for correctly inferred items; still, participants' MG responses to items incorrectly inferred in trial-and-error learning followed by definitions were as accurate as responses to correctly inferred items learned under errorless conditions. Inference errors were most detrimental to the development of declarative knowledge when they were not followed by definition, that is, when they remained uncorrected. Inference accuracy was not a reliable predictor of participants' reading times on the posttests (operationalized as a measure of non-declarative knowledge). SPR results also suggest that, both immediately after the learning treatment and two weeks later, the meanings of items experienced in the trial-and-error treatment were easier to integrate during reading compared to those in the errorless treatment.

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

In Experiment 2, we investigate whether these finding hold for L2 readers, whose lexical-semantic representations may be less robust than in the L1, particularly at lower L2 proficiencies (Elgort, Perfetti, Rickles, & Stafura, 2015; Favreau & Segalowitz, 1983; Frenck-Mestre & Prince, 1997; Phillips, Segalowitz, O'Brien, & Yamasakia, 2004) and for learners who acquired the L2 later in life (Silverberg & Samuel, 2004). As a consequence of having weaker lexical-semantic networks, L2 readers may be less able to predict and infer meanings of unfamiliar words in reading, whether overtly or covertly. This could create an advantage for errorless over trial-and-error learning because presenting a definition upfront supports the encoding of correct semantic features of a novel word. However, this potential errorless learning advantage may not persist if inference attempts are followed by definitions because conceptual errors facilitate retrieval (Cyr & Anderson, 2015) and encourage deeper engagement with definitions (Potts et al., 2018).

EXPERIMENT 2

METHODOLOGY

In Experiment 2 conducted with L2 participants, the materials, design, posttests, and procedures were the same as in Experiment 1. In the MG posttest, in addition to explanations and synonyms in English, participants could use their L1, Chinese. Chinese responses were translated by a balanced Chinese-English bilingual not involved in the study.

Participants

Study participants were 52 university students (41 female), who self-identified as native speakers of Chinese and second language speakers of English. Their average age was 22.6 (SD = 3.05). Their average LexTALE score was 72% (SD = 9.65%), that is, 22 percentile points lower than that of the L1 participants. Among them, seven were enrolled in the most advanced class of the university English Proficiency Programme, and the rest were undergraduate students: 17 were studying humanities and law and 27 science, engineering, or commerce. They received a $50 voucher for their participation.

RESULTS

Participants' responses were scored by two researchers, with an initial agreement of 93% for the inference accuracy and 98% for the MG accuracy.

Contextual Inferences during the Initial Learning Treatment

Across all treatments, the mean Inference accuracy was 63%. Responses shorter than one second were excluded from the analysis (33 data points; 0.70%). An average inference time was 20 seconds (SD = 8.10 seconds). Inference accuracy was significantly better for the ERL compared with CTRL and TAE treatments ($z = 11.20$ adjusted $p < .001$; $z = 10.23$ adjusted $p < .001$, respectively; Appendix C, Tables C.1a and C.1b). Participants

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

who had spent more time in English-speaking countries generated more accurate contextual meaning inferences ($z = 3.38$ $p < .001$). Slower inference generation times were observed on more accurate responses ($z = 6.99$ $p < .001$).

Meaning Generation

There was a statistically significant two-way interaction ($z = -2.52$ $p < .05$) in the analysis of MG scores between Treatment and Inference accuracy (Appendix C, Tables C.2a and C.2b). For correct meaning inferences (Figure 4), the MG accuracy for TAE items was significantly higher than for ERL ($z = 2.62$ adjusted $p < .05$) and numerically higher than for CTRL items ($z = 2.29$ adjusted $p = .11$). For incorrect meaning inferences, the MG accuracy for TAE items was significantly higher than for CTRL items ($z = 4.22$ adjusted $p < .01$) and numerically higher than for ERL items ($z = 2.44$ adjusted $p = .07$). The MG accuracy for ERL items was not significantly different from CTRL items, regardless of the inference accuracy.

There were also main effects of Treatment, Inference accuracy, and Session (immediate/delayed) in this analysis. The odds of an accurate meaning generation for TAE items were 2.86 times the odds for CTRL items (95% CIs: 1.76,4.67; $d = .58$, i.e., greater than .4), but the accuracy of responses to ERL items was not significantly different from CTRL and the effect size of this comparison was small ($d = .14$). Inference accuracy was a significant predictor of the MG accuracy ($z = 5.53$ $p < .001$ $d = .56$), with about 2% MG accuracy improvement when contextual meaning inferences were correct. The

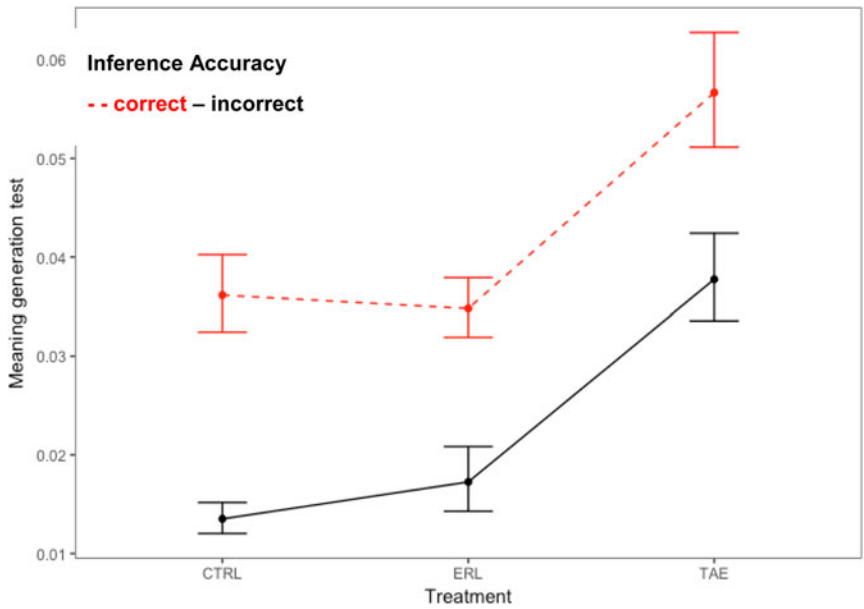


FIGURE 4. The estimated interaction between *Treatment* and *Inference accuracy* (Guess.ACC) in the analysis of L2 MG scores (fit and 95% pointwise confidence intervals). *Note*: CTRL = CWL without definitions, ERL = errorless treatment; TAE = trial-and-error treatment followed by definitions.

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

participants' MG response accuracy was about 5% higher on the immediate than the delayed posttest ($z = 10.70$ $p < .001$ $d = .86$).

We also compared meaning generation scores for items correctly inferred (encoded) in the ERL treatment with items correctly inferred in the TAE treatment (TAE.ACC.1) and items incorrectly inferred in the TAE treatment (TAE.ACC.0) (Figure 5). Meaning generation for correctly inferred TAE items was significantly more accurate compared with correctly inferred ERL items ($z = 3.81$ adjusted $p < .001$) and incorrectly inferred TAE items ($z = 2.40$ adjusted $p = .04$); correctly inferred ERL items had no advantage over incorrectly inferred TAE items ($z = 1.24$ adjusted $p = .43$).

Self-Paced Reading

The response accuracy to the comprehension questions in the SPR task was 80.66%; it was 82% on the immediate SPR and 79% on the delayed SPR. One participant's data were excluded from the SPR analysis due to a close-to-chance comprehension accuracy score.

Immediate SPR

After trimming extreme outliers (19 data points, 0.41%), the mean RT on the target was 531 ms (392 ms); and the mean RT on the $n + 1$ was 463 ms (251 ms). There was no reliable effect of Treatment or Inference accuracy in the immediate SPR. RTs on the target and $n + 1$ were faster for the health than building items ($t = -3.95$ $p < .001$; $t = -4.30$ $p < .001$, respectively). L2 readers were slower when reading the seven-letter than six-letter targets ($t = 2.00$ $p < .05$). Comparisons of RTs on the target and $n + 1$ for correctly inferred (encoded) ERL items with correctly and incorrectly inferred TAE items showed no significant differences.

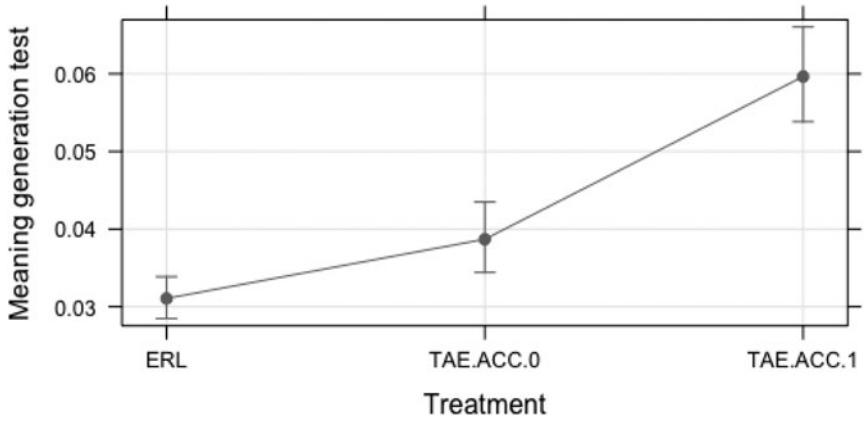


FIGURE 5. The estimated results for the correctly inferred items encountered in the errorless treatment (ERL), correctly inferred trial-and-error items (TAE.ACC.1), and incorrectly inferred trial-and-error items (TAE.ACC.0) in the analysis of L2 MG scores (fit and 95% pointwise confidence intervals).

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

Delayed SPR

After trimming extreme outliers (14 data points, 0.31%), the mean RT on the target was 395 ms (SD = 238), that is, 135 ms faster than on the immediate SPR; and it was 374 ms (SD = 184) on the $n + 1$, that is, 89 ms faster than on the immediate SPR. The analysis of target RTs did not show significant differences between the three treatments.

In the $n + 1$ analysis, RTs were predicted by Treatment, RTs on the preceding target, Accuracy of comprehension and Vocabulary scores but not Inference accuracy (Appendix C, Table C.3a and C.3b). RTs on $n + 1$ were faster for TAE than CTRL items ($t = -2.79$ adjusted $p = .014$) but the difference was very small (7.1 ms) and so was the effect ($d = .07$). RTs on $n + 1$ for TAE items were also 5.2 ms faster than for ERL items but this contrast did not reach statistical significance ($t = -2.06$ adjusted $p = .097$). Reading times on $n + 1$ were not different for the ERL and CTRL treatments. RTs on $n + 1$ were faster for participants whose comprehension scores were higher and whose L2 vocabulary test scores were higher; however, the effects of these individual predictors became nonsignificant after the p value adjustment for false discovery (Appendix C, Table C.3a). RTs on $n + 1$ were faster when RTs on preceding targets were faster ($t = 6.47$ $p < .001$).

RTs on correctly inferred (encoded) ERL items, correctly inferred TAE items, and incorrectly inferred TAE items were not significantly different. There was no significant difference in $n + 1$ RTs between correctly inferred ERL items and incorrectly inferred TAE items but there was a trend for $n + 1$ RTs to be faster for correctly inferred TAE items than for correctly inferred ERL items ($t = -1.98$ adjusted $p = .116$ $d = .06$). Other predictors of $n + 1$ RTs were RTs on the preceding target ($t = 6.76$ $p < .001$ $d = .30$), accuracy of text comprehension ($t = -2.68$ adjusted $p < .05$, $d = .11$), and vocabulary test scores ($t = -1.97$ adjusted $p = .131$, $d = .57$). There was no effect of Inference accuracy in the delayed SPR analyses.

DISCUSSION

Experiment 2 mostly replicated the results of Experiment 1, for L2 readers. Presenting a definition after a contextual encounter with a novel word produced better knowledge of meaning than presenting it prior to a contextual encounter. The errorless learning treatment did not differ on either nondeclarative or declarative knowledge measures from contextual exposure to the word without definitions. The accuracy of explicit meaning inferences in the learning treatment affected the scores on the meaning generation posttest but not the reading times in the SPR task. Time spent in English-speaking countries was predictive of participants' ability to derive word meanings from context. Participants' reading comprehension skills predicted their ability to integrate word meanings into text during reading on the delayed SPR, but no interaction between these proficiency-related variables and treatment was observed; this is counter to our prediction that an errorless learning advantage may be observed with less proficient L2 readers.

GENERAL DISCUSSION

The answer to our main research question is straightforward. Trial-and-error learning with definitions is more beneficial for the learning of meaning in L1 and L2 CWL than

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

errorless learning, for healthy adult participants. No advantage was observed for errorless over trial-and-error learning (with or without definitions) on the posttests of declarative knowledge. On the measure of nondeclarative knowledge (SPR), the advantage of the trial-and-error treatment followed by definitions was less robust. Although the items encountered in the trial-and-error treatment were integrated into the reading context significantly faster than those in the control treatment and the items encountered in the errorless treatment were not (as shown by the RTs on the spillover word), there was no statistically significant difference between the two treatment conditions (only a numerical one). In addition, the difference of 7 ms is very small (considering that RTs were in the range of 270 and 460 ms, across the two experiments), and the effect sizes were also small, even for *lme* analyses. Still, because all results are pointing in the same direction, that is, that RTs on the spillover word were faster for the trial-and-error treatment with definitions than the other two treatments, the hypothesis that the meanings of the novel words experienced in this treatment are easier to integrate into sentence contexts is worth exploring further. Based on the present results, future studies may benefit from including a higher number of participants per condition for the SPR measure; hopefully, future studies will also be in a better position to estimate required sample sizes based on our results.

Acknowledging that incorrect contextual inferences were more detrimental to CWL without definitions (on the declarative knowledge measure), the added value of definitions was not the same for trial-and-error and errorless learning. So, how can the advantage of trial-and-error over errorless CWL be explained? It is possible that the nature of the learning treatments differentially affects the quality of processing during initial contextual encounters with unfamiliar words. Meeting a novel word form while reading is likely to result in an attentional shift to that form, involving a deeper, more precise processing of the form and an effort to integrate its meaning (Chaffin, 1997; Chaffin, Morris, & Seely, 2001; Godfroid, Boers, & Housen, 2013; Williams & Morris, 2004). Presenting definitions of unfamiliar words before reading (errorless treatment) may reduce readers' attention to the novel form during reading because it may appear more familiar (Koriat & Bjork, 2005; Yang, Potts, & Shanks, 2017). Errorless learning may also limit the perceived need for making contextual inferences and limit their range, negatively affecting the quality of learning by reducing readers' engagement with contextual cues. Correct contextual meaning inferences in trial-and-error learning, however, are associated with deeper encoding (Rodríguez-Fornells et al., 2004).

Exposure to definitions before reading also reduces the likelihood of beneficial conceptual errors (i.e., errors that overlap semantically with correct answers), as evidenced by lower inference error rates in the errorless treatment in our study. Initial meaning inferences in supportive contexts, even when imprecise, are contextually motivated and may be integrated into the memory trace due to their semantic association with the target word, acting as "stepping stones" toward the target and serving as memory cues when explicit recall of word meanings is required (Cyr & Anderson, 2015). This is exactly what was observed in the meaning generation task. Additionally, covert meaning inferences for a novel word in supportive contexts are likely to be broadly aligned with the local and global meaning of the text (if it is understood) and contain features that would resonate with the core semantic features subsequently presented in a definition (Bolger et al., 2008). For example, an inference "a lofty tool that takes strength to wield"

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

submitted for the target, *adimper*, meaning *wrench*, was in the right conceptual category (tool) and had semantic features overlapping with the correct meaning (a hand tool). These overlapping features would likely resonate with the definition ("a hand tool that is used to hold or twist a nut or bolt"). This sequence of events facilitates the establishment of more robust meaning representations, which would explain the comparative ease of integration of the trial-and-error items into sentence contexts in SPR. Additionally, initial erroneous inferences may result in more cognitive activity during postretrieval error-checking, facilitating retention (Rodríguez-Fornells et al., 2004).

In considering the consequences of incorrect meaning inferences for CWL, we found a dissociation between declarative and nondeclarative knowledge. Inference errors negatively affected accuracy of meaning generation (particularly when they were not followed by definitions) but had no direct bearing on reading times in SPR, irrespective of whether definitions had been presented during the learning treatment. We therefore submit that explicit inferences may not be indicative of how semantic features are implicitly encoded during CWL. However, because nondeclarative knowledge takes longer to develop (Reber, 2013; Ullman & Lovelett, 2018), the impact of inference accuracy on the development of nondeclarative knowledge may be more subtle and difficult to detect in early CWL. The effect of incorrect contextual inferences on nondeclarative, implicit knowledge needs to be further tested using more sensitive instruments, such as eye-tracking, masked semantic priming, and electroencephalography (ERP).

L1 AND L2 READERS

Although the patterns of meaning inferences for the three treatments were very similar for L1 and L2 readers, there was a substantial difference in the absolute inference accuracy for the two groups; across all treatments, the mean inference accuracy of the L1 readers was 88% but it was 63% for L2 readers and, even in the ERL treatment, L2 readers' mean inference accuracy was more than 10% lower than that of L1 readers. The gap between inference accuracy in the errorless and trial-and-error treatments was twice as great for L2 than L1 participants. This suggests that L2 readers were less able to infer correct meanings of unfamiliar words from context (with only just more than half of the meanings inferred correctly in the control and trial-and error treatments) and that they were much more dependent on the upfront definitional support, even when explicitly instructed to derive word meanings from context. For L1 readers, a larger difference in the inference accuracy between the errorless and trial-and-error treatments was observed in a less familiar theme (building) than in a more familiar theme (health). Taken together these results suggest that participants' ability to explicitly derive word meanings from context decreases as text comprehension difficulty increases and topic familiarity decreases. Proportionally, vocabulary knowledge gains were an order of magnitude greater in the L1 than in the L2. Even when contextual inferences had been correct, the model predictions showed meaning generation accuracy of 20 to 30% for L1 readers but only 6% for L2 reader. Still, because it took participants longer to enter a correct than an incorrect inference in both experiments, we can be confident that the treatment effect was not due to item difficulty per se. Finally, the effect of attrition over time was greater for L2 than L1 readers (d = .86 compared to d = .45, respectively).

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

The trajectory of the nondeclarative knowledge acquisition also differed somewhat for L1 and L2 readers. The advantage of trial-and-error learning with definitions over errorless learning was observed in the immediate posttest in Experiment 1 and delayed posttest in Experiment 2, and it was more reliable in the L1 than L2 analyses. Still, when only correctly inferred items were compared, items experienced in the trial-and-error treatment were easier to integrate into sentence context than items from errorless treatment, for both L1 and L2 readers. Finally, only L1 readers showed an advantage for incorrectly inferred items in the trial-and-error treatment followed by definitions, compared with correctly inferred errorless items, in the delayed SPR. These findings suggest that L2 readers may be less effective than L1 readers in developing nondeclarative semantic knowledge from encountering novel words in supportive contexts. Our findings add to the evidence that L1 (and advanced L2) learners establish more robust lexical-semantic representations than less proficient L2 learners (Elgort et al., 2015; Favreau & Segalowitz, 1983; Frenck-Mestre & Prince, 1997; Phillips et al., 2004).

EDUCATIONAL IMPLICATIONS

Our study shows that vocabulary knowledge develops faster and is more robust when CWL is supplemented by dictionary-style definitions. Therefore, the use of dictionaries should be encouraged in conjunction with reading and not discouraged, as is the case in some university preparation programs for L2 students. However, students should be encouraged to infer meanings of unfamiliar words from context prior to looking them up. Textbook publishers, course material developers and university teachers should note that, although presenting definitions of terms and concepts before their appearance in the text may have initial benefits during reading (particularly, for L2 readers), if the goal is to increase and enhance students' specialized vocabulary knowledge, then providing definitions after reading is preferable for L1 and L2 readers. For students (particularly, L2 speakers), it could be reassuring to know that, as long as they verify their contextual guesses soon after reading, a negative effect of such guesses is minimal.

CONCLUSIONS

We conducted two experiments with L1 and L2 readers, under ecologically valid learning conditions, to establish which approach to CWL leads to more robust declarative and nondeclarative knowledge—errorless learning (with definitions made available before reading) or trial-and-error learning (with definition provided after reading). Research into errorless learning (i.e., reducing errors during encoding) has been inconclusive, with both errorless and trial-and-error learning producing better results with different participant populations, for different learning tasks and procedures, and on different measures (Clare & Jones, 2008; Middleton & Schwartz, 2012). Our study provides clear evidence in favor of trial-and-error contextual learning of word meanings for healthy adults. We have deliberated the reasons for the superiority of trial-and-error over errorless learning, considering contributions of both explicit and implicit learning to this process. We have also discussed the differential influence of explicit inferences made during reading on the development of declarative and nondeclarative word knowledge. Overall, this study makes a step toward a better understanding of how word knowledge

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

develops during and from reading, the role of definitions in contextual word learning, and the impact of inferential errors on declarative and nondeclarative knowledge of meaning.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263119000561

NOTES

[1]The label *nondeclarative* is used in our study, following Reber (2008, 2013), instead of *procedural* or *implicit*, to avoid controversies that are not central to the present study. Based on a review of neuropsychological and neuroimaging research showing a dissociation between memory systems that are dependent on the medial temporal lobe (MTL) and those that are not, Reber proposes the following definitions: "*declarative* memory is whatever memory processes are accomplished by the MTL memory system and everything else is *nondeclarative*" (Reber, 2013, p. 2026). Ullman and Lovelett (2018) argue that, although both memory systems can acquire analogous knowledge and skills, explicit knowledge (i.e., semantic knowledge of facts and episodic memory of personal events) is underpinned exclusively by declarative memory, and implicit knowledge (i.e., nonconscious knowledge of various sorts, resulting from implicit learning, such as repeated exposure, and supporting a wide range of perceptual-motor and cognitive skills) is underpinned exclusively by nondeclarative memory.

[2]Seabrooke, Hollins, Kent, Wills, and Mitchella (2019) also found that trial-and-error learning improved the recognition of the novel words and their definitions, however, it was less effective than errorless learning in creating word-meaning associations (Exp. 5). However, the development of form-meaning associations in the trial-and-error condition could have been disadvantaged in the experiment of Seabrooke et al. due to longer presentation times of the cue (meaning) in the errorless compared to trial-and-error learning condition.

[3]The term *errorless* refers to the type of treatment that does not require participants to generate a guess about a word's meaning prior to accessing its definition. However, this does not guarantee that no errors are made when the word is encountered in reading or when an explicit contextual inference is generated.

[4]To determine an appropriate sample size for this study, we ran a power simulation (aiming to achieve 80% power) using *SIMR* package (Green & MacLeod, 2016) in R (version 3.4.4; R Core Team, 2018). We used the meaning generation task data and models fitted in *lme4* from Elgort (2017) because this study posed a similar research question, measured declarative knowledge using the same test and was conducted with L2 participants of a similar background and proficiency to those in Experiment 2 in the present study. Unfortunately, we could not locate any studies that evaluated the development of nondeclarative knowledge in CWL using SPR; so, it was not possible to use the same procedure for the SPR measure. However, as recommended by Brysbaert and Stevens (2018, p. 16), we aimed to have about 1,600 observations per condition (ERL, TAE, and CTRL) for the SPR analysis.

[5]Word family (Bauer & Nation, 1993) is a unit of counting used in word frequency baselists organized in levels (or bands) of 1,000 base words. A word family consists of a base word (e.g., *cancel*) and its inflected forms (*cancelled*, *cancelling*), as well as derivative forms that share a common meaning with the base word (*cancellation*, *cancellations*). Baselists are used to measure the vocabulary load of reading texts and to estimate receptive vocabulary knowledge of languages users.

[6]Because we were interested in how access to definitions affects *initial* stages of CWL, three contextual occurrences appear to be sufficient to trigger the learning of meaning in supportive contexts, with both declarative and nondeclarative knowledge observed on the third occurrence (e.g., Mestres-Missé, Rodríguez-Fornells, & Münte, 2007).

[7]Brysbaert and Stevens (2018) explain why, in experiments with multiple observations per participant per condition, effect sizes reported in *lme* analyses are much smaller than those reported in traditional F1 and F2 analyses of variance (ANOVAs). This is because averaging observations over participants and items in F1 and F2 analyses reduces the variance in the data, thus reducing the standard error. Effect sizes calculated in *lme* models are smaller because the standard deviation is calculated across a large number of individual observations, generating a much higher level of noise. Brysbaert and Stevens (2018, pp. 4–6), for example, calculated

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

effect sizes for Adelman et al. (2014) orthographic priming study with 1,020 participants who made lexical decisions to 420 words. The traditional F1 and F2 ANOVAs showed large effect sizes ($d > .8$) of priming. However, in the *lme* reanalysis of the same data (using the approach of Westfall et al. 2014), the effect size was $d = .0868$.

REFERENCES

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., … Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*, 1052–1067. https://doi.org/10.3758/s13428-013-0442-y

Avery, N., & Marsden, E. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, 1–33. doi: 10.1017/S0272263119000196

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68. http://dx.doi.org/10.1016/0028-3932(94)90068-X

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279.

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*, 15–32. https://doi.org/10.1111/j.1467-9817.1984.tb00252.x

Bolger, D. J., Balass, M., Landen, E. & Perfetti, C. A. (2008). Contextual variation and definitions in learning the meaning of words. *Discourse Processes*, *45*, 122–159. https://doi.org/10.1080/01638530701792826

Bordag, D., Kirschenbaum, A., Tschirner, E., & Opitz, A. (2015). Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. *Bilingualism: Language and Cognition*, *18*, 372–390. https://doi.org/10.1017/S1366728914000078

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, 9. http://doi.org/10.5334/joc.10

Carpenter, S. K., Sachs, R. E., Martin, B., Schmidt, K., & Looft, R. (2012). Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin & Review*, *19*, 81–86. https://doi.org/10.3758/s13423-011-0185-7

Chaffin, R. (1997). Associations to unfamiliar words: Learning the meaning of new words. *Memory and Cognition*, *25*, 203–226. https://doi.org/10.3758/BF03201113

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 225–235. http://dx.doi.org/10.1037/0278-7393.27.1.225

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*, 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review*, *18*, 1–23. https://doi.org/10.1007/s11065-008-9051-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 841–850. http://dx.doi.org/10.1037/xlm0000073

Cyr, A.-A., & Anderson, N. D. (2018). Learning from your mistakes: Does it matter if you're out in left foot, I mean field? *Memory*, *26*, 1281–1290. http://doi.org/10.1080/09658211.2018.1464189

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

Dóczi, B., & Kormos, J. (2016). Second language learning and vocabulary development. In B. Dóczi & J. Kormos, *Longitudinal developments in vocabulary knowledge and lexical organization* (pp. 118–138). New York: Oxford University Press.

Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, *1*, 1–11. http://doi.org/10.22599/jesla.3

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*, 341–366. https://doi.org/10.1017/S0272263117000109

Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J. Z. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience*, *30*, 506–528. http://dx.doi.org/10.1080/23273798.2014.942673

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, *64*, 365–414. https://doi.org/10.1111/lang.12052

Favreau, M., & Segalowitz, N. (1983). Automatic and controlled processes in the first and second language reading of fluent bilinguals. *Memory and Cognition*, *11*, 565–574. https://doi-org.helicon.vuw.ac.nz/10.3758/BF03198281

Frantzen, D. (2003). Factors affecting how second language Spanish students derive meaning from context. *Modern Language Journal*, *87*, 168–199. https://doi.org/10.1111/1540-4781.00185

Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, *21*, 225–241.

Frenck-Mestre, C., & Prince, P. (1997). Second language autonomy. *Journal of Memory and Language*, *37*, 481–501. https://doi.org/10.1006/jmla.1997.2526

Frishkoff, G. A., Collins-Thompson, K., Hodges, L. L., & Crossley, S. (2016). Accuracy feedback improves word learning from context: Evidence from a meaning-generation task. *Reading and Writing*, *29*, 609–632. https://doi.org/10.1007/s11145-015-9615-7

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., … Yoon, H. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, *21*, 563–584. https://doi.org/10.1017/S136672891700021.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition*, *35*, 483–517. https://doi.org/10.1017/S0272263113000119

Green, P., & MacLeod, C. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498.

Haist, F., Shimamura, A. P., & Squire, L. R. (1992). On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 691–702. http://dx.doi.org/10.1037/0278-7393.18.4.691

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal 50*, 346–363. https://doi.org/10.1002/bimj.200810425

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*, 514–527. https://doi.org/10.3758/s13421-011-0167-z

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). London, UK: Macmillan. https://doi.org/10.1007/978-1-349-12396-4_11

Hulstijn, J. H. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research*, *18*, 193–223. https://doi.org/10.1191/0267658302sr207oa

Jenkins, J. R., Stein, M., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, *21*, 767–787. http://dx.doi.org/10.2307/1163000

Joseph, H. S. S. L., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, *133*, 238–248. https://doi.org/10.1016/j.cognition.2014.06.015

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238. http://dx.doi.org/10.1037/0096-3445.111.2.228

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

Karpicke, J., & Roediger, H. L. III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*, 704–719. http://dx.doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. http://dx.doi.org/10.1126/science.1152408

Kelly, P. (1990). Guessing: No substitute for systematic learning of lexis. *System*, *18*, 199–207. https://doi.org/10.1016/0346-251X(90)90054-9

Kim, M., Crossley, S. A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading & Writing*, *31*, 1155–1180. https://doi.org/10.1007/s11145-018-9833-x

Knight, J. B., Ball, H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*, 731–746. http://dx.doi.org/10.1016/j.jml.2011.12.008

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 187–194. http://dx.doi.org/10.1037/0278-7393.31.2.187

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.) *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge, UK: Cambridge University Press.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*, 89–108. https://doi.org/10.1177/003368829702800106

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*, 861–904. https://doi.org/10.1017/S0142716418000036

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

Medler, D. A., & Binder, J. R. (2005). *MCWord: An on-line orthographic database of English language*. Retrieved from http://www.neuro.mcw.edu/mcword/

Mestres-Missé, A., Rodríguez-Fornells, A., & Münte, T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, *17*, 1858–1866.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*, 225–229. https://doi.org/10.3758/BF03194056

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation*, *22*, 138–168. https://doi.org/10.1080/09602011.2011.639619

Mondria, J. (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*, 473–499. https://doi.org/10.1017/S0272263103000202

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Paradis, M. (1994). Neurolinguistic aspects of implicit and explicit memory: Implications for bilingualism and SLA. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 393–419). London, UK: Academic Press.

Pellicer-Sánchez, A. (2015). Incidental L2 vocabulary acquisition *from* and *while* reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*, 97–130. https://doi.org/10.1017/S0272263115000224

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*, 31–55.

Perfetti, C. A., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22–37. https://doi.org/10.1080/10888438.2013.827687

Phillips, N. A., Segalowitz, N., O'Brien, I., & Yamasakia, N. (2004). Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics*, *17*, 237–262. https://doi.org/10.1016/S0911-6044(03)00055-1

Potts, R., Davies, G., & Shanks, D. R. (2018, July 19). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 1023–1041. http://dx.doi.org/10.1037/xlm0000637

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. http://dx.doi.org/10.1037/a0033194

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Reber, P. J. (2008). Cognitive neuroscience of declarative and nondeclarative memory. In Aaron, A. S. Benjamin, J. S. De Belle, B. Etnyre, & T. A. Polk (Eds.), *Advances in psychology*, Vol. *139* (pp. 113–123). Amsterdam and New York: North-Holland Publishing Company.

Reber, P. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*, 2026–2042. https://doi.org/10.1016/j.neuropsychologia.2013.06.019

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading*, *7*, 219–237. https://doi.org/10.1207/S1532799XSSR0703_2

Rieder, A. (2002). A cognitive view of incidental vocabulary acquisition: From text meaning to word meaning. *VIEWS*, *11*, 53–71. Retrieved from https://www.univie.ac.at/Anglistik/views/02_1&2/AR.PDF

Rodríguez-Fornells, A., Kofidis, C., & Muente, T. F. (2004). An electrophysiological study of errorless learning. *Cognitive Brain Research*, *19*, 160–173. https://doi.org/10.1016/j.cogbrainres.2003.11.009

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*, 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime (Version 2.0) [Computer software and manual]. Pittsburgh, PA: Psychology Software Tools Inc.

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82. https://doi.org/10.1016/j.jml.2018.10.001

Silverberg, S., & Samuel, A. G. (2004). The effect of age of second language acquisition on the representation and processing of second language words. *Journal of Memory and Language*, *51*, 381–398. https://doi.org/10.1016/j.jml.2004.05.003

Stahl, S., & Fairbanks, M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*, 72–110. https://doi.org/10.3102/00346543056001072

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *34*, 39–65. https://doi.org/10.1177/0267658316675195

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.

Warmington, M., & Hitch, G. J. (2014). Enhancing the learning of new words using an errorless learning procedure: Evidence from typical adults. *Memory*, *22*, 582– 94. https://doi.org/10.1080/09658211.2013.807841

Warmington, M., Hitch, G. J., & Gathercole, S. E. (2013). Improving word learning in children using an errorless technique. *Journal of Experimental Child Psychology*, *114*, 456–465. https://doi.org/10.1016/j.jecp.2012.10.007

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. https://doi.org/10.1093/applin/aml048

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

Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*, 312–339. https://doi.org/10.1080/09541440340000196

Witzel, N., Witzel, J. & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, *41*, 105–128. https://doi-org.helicon.vuw.ac.nz/10.1007/s10936-011-9179-x

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1073–1092. http://dx.doi.org/10.1037/xlm0000363

https://doi.org/10.1017/S0272263119000561 Published online by Cambridge University Press