Nakata, T. & Elgort, I. (2020) Accepted MS for the Special Issue of SLR
"Lexical acquisition and processing: Setting an interdisciplinary research agenda".

Effects of Spacing on Contextual Vocabulary Learning: Spacing Facilitates the Acquisition of Explicit, But Not Tacit, Vocabulary Knowledge

Abstract

Studies examining decontextualized associative vocabulary learning have shown that long spacing between encounters with an item facilitates learning more than short or no spacing, a phenomenon known as distributed practice effect. However, the effect of spacing on learning words in context is less researched and the results, so far, are inconsistent. In this study, we compared the effect of massed and spaced distributions on second language vocabulary learning from reading. Japanese speakers of English encountered 48 novel vocabulary items embedded in informative English sentences, inferred their meanings from contexts, and received feedback in the form of English synonyms and Japanese translation equivalents. To test the hypothesis that the effects of spacing might differentially affect the development of explicit or tacit word knowledge, spacing effects were measured using semantic priming as well as a meaning recall and a meaning-form matching posttest. Results showed an advantage of spaced over massed learning on the meaning recall and meaning-form matching posttests. However, a similar semantic priming effect was observed irrespective of whether an item was encountered in the massed or spaced distribution. These results suggest that the spacing effect holds in contextual word learning for the development of explicit vocabulary knowledge, but massing appears to be as effective as spacing for the acquisition of tacit semantic knowledge.

Accepted version

Research on the effects of spacing in learning has a long history dating back to the 19th century (Ebbinghaus, 1885). In the majority of studies, researchers have followed the tradition of verbal memory research and examined the effects of spacing on paired-associate learning where participants are asked to memorize target words and their meanings (i.e., associates). In second language (L2) paired-associate word learning studies, the effect of spacing has been mostly positive (e.g., Bahrick and Phelps, 1987; Karpicke and Bauernschmidt, 2011; Nakata, 2015; Nakata and Suzuki, 2019); introducing long spacing between encounters with a given item facilitates learning more than short or no spacing, a phenomenon known as distributed practice effect. The effect of spacing in more naturalistic, contextual vocabulary learning, such as learning vocabulary from reading, has been less researched, and recent L2 contextual word learning studies have yielded variable results (Elgort et al., 2018; Elgort and Warren, 2014; Koval, 2019; Serrano and Huang, 2018; Webb and Chang, 2015). The aim of the present study was to compare the effects of massed and spaced distributions on incidental L2 vocabulary learning from context followed by feedback. Because explicit and tacit word knowledge may be differentially affected by the spacing of contextual encounters and the inference-feedback cycles, we used posttests that measured both explicit knowledge (meaning-form matching and cued meaning recall) and tacit knowledge (semantic priming task).

Effects of Spacing on L2 Vocabulary Learning

When discussing the effects of spacing, it is useful to make a distinction between the spacing effect and lag effect (Rogers, 2017). The spacing effect refers to the phenomenon where spaced learning, which involves spacing between encounters with a given item, yields superior retention relative to massed learning, which does not involve any spacing (Cepeda et al., 2006). The lag effect, in contrast, refers to the phenomenon where longer spacing generally leads to better long-term retention than shorter spacing (Cepeda et al., 2006; Rogers, 2017). In other words, while the spacing effect is concerned with the effects of spacing and no spacing (i.e., massing), the lag effect pertains to the effects of different amounts of spacing. The term *distributed practice effect* is sometimes used to collectively refer to both spacing and lag effects.[1] The spacing effect is considered "one of the most

Accepted version

robust phenomena in experimental psychology" (Ellis, 1995, p. 118). Nakata (2015) compared the effects of massed and spaced schedules on L2 vocabulary learning. In the massed condition, target items were repeated four times in a row. In the spaced condition, the target words were also repeated four times, but after 30 intervening trials (approximately 6 minutes). Posttests conducted a week after the treatment showed that spaced learning was more than twice as effective as massed learning. The spacing effect has also been observed in other L2 vocabulary studies conducted in the paired-associate paradigm (e.g., Karpicke and Bauernschmidt, 2011; Seibert, 1932).

While the findings yielded by extant L2 vocabulary research are very valuable, most earlier spacing studies are limited to the paired-associate learning paradigm, in which participants were explicitly asked to memorize the target words and their meanings. Although empirical evidence suggests that paired-associate learning is effective, efficient, and useful (Elgort, 2011), it should constitute only a small part of the language learning curriculum dominated by meaning-focused learning, as suggested in the four strands approach (Nation, 2013). Given the importance of meaning-focused input for vocabulary development, it is surprising how few L2 vocabulary studies have investigated whether the spacing effect (i.e., the superiority of spaced over massed repetition) holds in contextual learning, for example, during reading (Koval, 2019). Laufer (2003) and Nation (2013) argue that, in incidental L2 vocabulary learning during reading, new words should be encountered again soon after the initial encounter in the input before they are forgotten. If this is the case and episodic memory traces of encounters with a new word in reading decay relatively quickly, shorter intervals between contextual encounters may be more beneficial for learning new words than longer intervals. Findings reported by Elgort and Warren (2014), who investigated incidental L2 vocabulary learning from reading, support this conjecture.

Another reason why spacing between encounters may not enhance contextual vocabulary learning is that spacing might decrease inductive learning. As Kornell and Bjork (2008) observed, spacing is sometimes considered the "enemy of induction" (p. 585) because juxtaposing multiple exemplars of a given concept at once (massing) may enable learners to discover the features that define a particular concept, potentially making induction easier.

Accepted version

Spacing, in contrast, may make induction more difficult because long intervals potentially limit the learners' ability to notice similarities between exemplars. In decontextualized vocabulary learning, learners are typically given the meaning of target words at the outset, so there is no need to use induction. In incidental contextual word learning, learners have to infer meanings of unfamiliar words from context, which is more likely to rely on inductive learning processes. If spacing reduces opportunities for and/or success of inductive learning in contextual vocabulary learning, it may negatively affect readers' ability to acquire new vocabulary from context. For instance, when learners are presented with multiple sentences containing a new word in a massed (rather than spaced) schedule, they may have more contextual clues available at once to infer its meanings. This may lead to more successful meaning inferences which, in turn, could lead to larger vocabulary gains. However, cognitive psychology research results regarding the effect of spacing on inductive learning are somewhat inconsistent (e.g., Kornell and Bjork, 2008; Kurtz and Hovland, 1956; Zulkiply and Burt, 2013) and, therefore, any strong hypotheses regarding the effect of spacing on inductive learning would be premature.

Conversely, L2 vocabulary research that is concerned with the effect of testing and retrieval (e.g., Barcroft, 2007; Karpicke & Roediger, 2008; van den Broek et al., 2018; Barcroft, 2015a) suggests that contextual vocabulary learning may benefit more from a spaced than massed repetition schedule. Retrieval is defined as the process of recalling previously learned information. For instance, when a new word is encountered in context following spaced distribution, each subsequent encounter after the first one represents an opportunity to retrieve stored information about the word from previous encounters. In contrast, multiple instances of the same new word close together (massed distribution) are likely to be encoded as one learning episode; thus, retrieval from previous learning episodes is not possible. In addition, this positive effect of retrieval on learning might be more pronounced when the correct meaning is provided as feedback after the inference attempt because it helps learners to verify and encode correct form-meaning connections (e.g., Carpenter et al., 2012; Metcalfe and Kornell, 2007).

Accepted version

Effects of Spacing on Contextual Vocabulary Learning

Recent studies that have investigated the effects of spacing on contextual L2 vocabulary learning in long continuous texts produced inconsistent results. In Serrano and Huang (2018), 71 Taiwanese high school students were divided into intensive and spaced distribution groups and read and listened to an English text (419 words) five times. While the intensive group read the same text on 5 consecutive days, the spaced group read it once per week over a 5-week period. Posttest results on a bilingual matching test showed that, while the intensive group obtained higher scores in the short term, the spaced group retained more words during the period between the immediate and delayed posttests. No significant difference, however, was found for gains from the pretest to the delayed posttest between the two groups.

In the study conducted by Elgort et al. (2018), 40 Dutch-speaking students read chapters from a nonfiction English book (12,152 words). The text was divided into two parts of approximately the same length, which the participants read over 2 days. The main research questions of this study were to do with incremental learning of new L2 words from reading a long authentic text, tracing the development of different component representations of new words in real time. Learning was measured by online (eye-tracking) and off-line (meaning recall) measures. The text contained 14 low-frequency target words, four of which occurred only on day one, five occurred only on day two, and five occurred on both days. This created an opportunity to compare the learning for words encountered only within the same day (shorter spacing) with that for words encountered across the 2 days (longer spacing). The meaning recall test results showed that encountering the target words across 2 days led to higher gains than encountering them within the same day. A similar result was observed for the eye-movement measure that indexes ease of integration of the word's meaning into the preceding context (*go-past time*) on the last occurrence of the word in the text. However, fewer regressions back to the target word (on the final encounter in the text) were observed when it was encountered on the same day. Elgort et al. hypothesized, therefore, that shorter (same-day) spacing may be more beneficial for the learning of form, and longer spacing (over 2 days) more beneficial for the learning of meaning. However, because the number of

Accepted version

occurrences of the target words across the short and long spacing were not deliberately manipulated or controlled, this hypothesis needs to be further tested.

Unlike Serrano and Huang (2018) and Elgort et al. (2018), Elgort and Warren (2014) and Webb and Chang (2015) have failed to find any benefits of long spacing in contextual vocabulary learning. In the study conducted by Elgort and Warren (2014), 48 adult ESL participants read four chapters from a nonfiction book (approximately 40,000 words) over 10 days, with 48 pseudowords that occurred multiple times in the text. Learning was measured by online (form and semantic priming) and off-line (meaning recall) measures. Results of the meaning recall task showed that the pseudowords repeated within the same chapter (shorter spacing) were acquired more successfully than those repeated across different chapters (longer spacing); moreover, the lower-proficiency participants were only able to contextually learn the pseudowords when they occurred within the same chapter. In Webb and Chang (2015), 61 Taiwanese secondary school students read and listened to 10 graded readers. Results from a bilingual form-meaning matching test revealed no statistically significant correlation between the distribution of occurrence (number of graded readers in which target words appeared) and vocabulary gains. The results reported by Elgort and Warren (2014) and Webb and Chang (2015) suggest that, the positive lag effect is not always observed in contextual vocabulary learning during reading long continuous texts and may vary with the lag duration. In these studies, however, spacing was not deliberately manipulated to investigate its effect on word learning, and their operationalization of shorter and longer spacing was opportunistic – based on the distribution of words in the authentic texts used as reading materials. Therefore, we are not really comparing apples with apples; that is, longer spacing in one study (e.g., encounters with a word over 2 days in Elgort et al., 2018) may be considered shorter spacing in another (e.g., in Elgort and Warren, some students took more than 1 day to complete a chapter). Importantly, none of the L2 contextual learning studies reviewed here compared massed (no spacing) and spaced repetition; instead they investigated the lag effect.

Massed and spaced repetition was compared in a recent contextual word learning study by Koval (2019). Twenty-four novel (Finish) words were encountered by 40 English-

Accepted version

speaking university students in sentence contexts, in a massed or spaced repetition schedule. The study used eye-tracking to investigate whether repetition affects attentional processing of novel L2 words differently under massed and spaced conditions, and whether the effect of spacing on word learning gains is mediated by attention. Koval found that the processing of the target words in the massed condition, that involved reading them in four consecutive sentences, was characterized by deficient attentional processing compared to the processing in the spaced condition. She also found a clear spacing effect on the immediate and delayed form recognition and form–meaning mapping posttests. This study suggests a potential advantage for spaced presentations in contextual word learning. However, Koval's study was interested in deliberate contextual learning (participants were pre-warned about the posttests of target words and instructed to try to memorize them during the learning phase). Moreover, the English meaning of the target word was presented prior to reading, which reduced the need to infer the meanings of the novel words from context. Therefore, this study does not address the question of whether the spacing effect is present in incidental contextual word learning. Also, knowledge gains were measured using off-line tests only. Therefore, the question of whether explicit and tacit knowledge, gained in incidental contextual word learning, is differentially affected by the massed and spaced presentation schedule, remains open.

The Present Study

The aim of the present study was to examine the effects of spacing on L2 vocabulary learning from reading. This study differs from earlier research in two major respects: (1) we investigate whether the spacing effect is present in incidental contextual word learning by deliberately manipulating massed (no spacing) and spaced schedules of encounters with novel L2 vocabulary, and (2) we assess whether the spacing effect is observed in explicit and tacit vocabulary knowledge. We hypothesized that the spacing effect may be present in the acquisition of explicit vocabulary knowledge but not necessarily in the acquisition of tacit knowledge of meaning. Studies that investigated the lag effect in the acquisition of grammar suggest that it may differentially affect different types of knowledge: long spacing may be more effective for the acquisition of explicit knowledge (e.g., measured using grammaticality

Accepted version

judgments; Bird, 2010), and short spacing may be more effective for the knowledge proceduralization, as measured by speed of sentence generation (Suzuki and DeKeyser, 2017). For contextual word learning, Elgort et al. (2018) found that longer spacing (repetition across 2 days) led to superior knowledge of meaning on both the off-line (meaning recall test) and the online eye-movement (go-past time) measures, but shorter spacing (encountering a word on the same day) resulted in faster online processing of the novel words' form during reading; and no effect of spacing was observed in neutral sentence contexts, on posttest. Elgort and Warren (2014) also failed to observe the lag effect on tacit word knowledge in their contextual word learning study. In the current study, we address the following research questions:

Research Question 1: Does spacing facilitate L2 vocabulary learning from context as compared with massed learning?

Research Question 2: Does spacing have differential effects on the acquisition of explicit and tacit vocabulary knowledge?

Before describing the detailed methodological information, we provide a brief overview of the current experiment. In this study, Japanese learners of English encountered 48 novel vocabulary items embedded in informative English sentences under two conditions: massed and spaced. In the massed schedule, participants read three sentences with the target pseudoword presented on the same screen (no lag), then inferred its meaning and then received feedback in the form of English synonyms and Japanese translation equivalents. In the spaced condition, the participants read each of the three sentences separately with a lag of about 25 minutes; they inferred the meaning of the pseudoword after each encounter and reviewed the correct meaning after each inference attempt.

To test the hypothesis that the effects of spacing might differentially affect the development of explicit or tacit word knowledge, we used immediate and delayed posttests that measured both explicit knowledge (meaning-form matching and cued meaning recall) and tacit knowledge (semantic priming combined with the lexical decision task). Although the two types of knowledge have been labelled and defined in SLA studies as explicit/implicit, declarative/procedural (or nondeclarative), more/less automatized, available

Accepted version

online/off-line; we use *explicit* and *tacit* to describe the type of knowledge tested in our study. This is because in the off-line tests of explicit knowledge participants can use all sorts of explicit decision and metacognitive strategies to select or generate an answer. There is no guarantee, however, that this kind of knowledge can be accessed in an online, non-effortful manner needed in fluent reading. In primed lexical decisions, the explicit decision made by the participants (i.e., whether a string of letters is a word) is *not* itself the measure of tacit word knowledge. We are interested in the difference between the reaction time (RT) to the same known L2 word target preceded by a semantically related versus unrelated prime (and in this study, related primes are incidentally learned pseudowords). Faster processing of the target in the related condition suggests that its lexical semantic representation has been preactivated by the prime overlapping with it in meaning. For this, the meaning of the prime had to be activated, even though the decision itself does not require the participants to fully access the meaning. This covert, tacit process is underpinned by participants' tacit knowledge of meaning. Tacit knowledge is also needed in fluent, low-effort access to contextually relevant word meanings during reading.

Method

Participants

The participants were 66 Japanese undergraduate and postgraduate students (44 of whom were female) with the average age of 21.9 (*SD* = 6.2). All but four participants were English majors. The remaining four participants were majoring in social sciences such as economy, sociology, and policy studies. All participants had normal or corrected-to-normal vision. The average English vocabulary size of the participants as measured by Vocabulary Size Test (hereafter, VST; Nation and Beglar, 2007) was 8,698.5 (*SD* = 1,136.6) word families, suggesting that they were higher-intermediate to advanced learners of English. The participants volunteered to participate and received 6,000 yen as compensation for their time.

Materials

Forty-eight orthographically and phonologically plausible pseudowords (6-7 letters) were used as target items (e.g., *bondit, emband, shottle*). Half of the pseudowords were related to the theme of building/household (e.g., *craftsman, shelf, detergent*), and the other

Accepted version

half were related to cooking/food (e.g., *spatula, menu, omelet*). Within each theme, care was taken not to use similar objects as referents (e.g., *knife, spoon, fork*) that could, when learned together, hinder learning. Instead, we used items that were related to either the schema of housing or to the schema of cooking, an approach that facilitates word learning (Tinkham, 1997). The pseudowords were divided into two sets of 24 items, Set A and Set B, containing 12 items from the two themes (building/household and cooking/food). In a counterbalanced within-participant design, half of the participants encountered Set A under the massed condition and Set B under the spaced condition, while it was reversed for the other half of the participants. Three sentences were prepared for each of the pseudowords, resulting in 144 sentences (48 pseudowords × 3 sentences). The sentences were based on those used by Elgort (2017), but were modified after piloting with eight Japanese college students. Specifically, for potentially problematic target items, more contextual clues were added to the sentences, or the meanings of the target items were simplified to facilitate successful meaning inference. The average length of the sentences was 15.3 words (*SD* = 3.8).

Posttest Design and Dependent Measures

We used three dependent measures of vocabulary knowledge: meaning recall, meaning-form matching, and semantic priming. For the meaning recall and meaning-form matching tests, response accuracy was used as dependent variables. In the lexical decision task, the dependent variables were response accuracy and RT on the targets. Priming effect was operationalized as the difference in RT in the related and unrelated condition. In our additional analysis, we used accuracy and RT of lexical decisions to the primes as dependent variables.

In the meaning recall posttest, participants were presented with pseudowords one at a time in isolation and asked to provide their meaning either in L1 (Japanese) or L2 (English). The test was given twice, in Session 1 (on the same day as the treatment) and Session 2 (2 days after the treatment). In the meaning-form matching posttest, the Japanese translation and English synonym of the pseudowords were presented one at a time, and participants were asked to choose the corresponding pseudoword from four options. The distractors used in the matching test were other pseudowords encountered during the treatment. To minimize the

Accepted version

effects on the other tests, the meaning-form matching posttest was administered only in Session 2.

In the tacit knowledge of meaning posttest, participants completed a lexical decision task that included a semantic priming manipulation. The pseudowords encountered by participants in the learning phase were paired with 48 English word targets related to them in meaning (e.g., the target word *shelter* was related in meaning to the pseudoword *emband*, meaning 小屋 - *shelter*; the target word *menu* was related to the pseudoword *ganset*, meaning メニュー - *menu*). The same word targets were also paired with unrelated word primes (e.g., *throat – shelter*; *nature – menu*). Two counterbalanced experimental lists were created, A and B; if the word target was preceded by an unrelated prime in List A, it was preceded by a related prime in List B; and vice versa. In each list, one half of the pseudoword primes was encountered by the participant in the massed condition during the treatment and the other half in the spaced condition. For the lexical decision task, 48 nonword targets were also added to the lists, half of which were preceded by word primes and half by nonword primes. Each list also contained 96 unrelated filler pairs, half of which were word primes paired with nonword targets and half nonword primes with word targets. Thus, each list also contained 192 pairs of stimuli; the proportion of related pairs was 12.5%. The related pairs were significantly more semantically similar (mean LSA index = 0.73) than the unrelated pairs (mean LSA index = 0.06; t (47) = 13.20, $p < .001$), as calculated using the Latent Semantic Analysis (LSA) tool (http://lsa.colorado.edu). Each list also included 50 practice trials. The task was administered twice, in Session 1 (on the same day as the treatment) and Session 2 (2 days after the treatment). To avoid the practice effect, participants who were assigned to list A in Session 1 received List B in Session 2, and vice versa.

Procedure

Both sessions were conducted with each participant individually in a quiet office of the first author. Session 1 consisted of the learning treatment and the immediate posttests. At the outset of Session 1, the participants received explanations about the study and signed a consent form. After that, the treatment was conducted using computer software developed for this study (written in Microsoft Visual Basic). During the treatment, participants were

Accepted version

presented with 144 sentences, containing one target pseudoword each. The participants were instructed to read the sentence fully, guess the meaning of the pseudoword and type their answer either in their L1 (Japanese) or L2 (English). In the massed condition, three sentences with the target pseudoword were presented simultaneously for 90 seconds as shown below (note that the pseudoword *emband* means *shelter* 小屋):

> The (emband) seems the ideal place to stay the night, if the storm continues.
> The (emband) was very old, its stone walls were broken, and the wind was whistling through the remains.
> We must walk faster if we want to reach the (emband) before dark.

In the spaced condition, only one sentence with the target pseudoword was presented at a time, and each following encounter with a given pseudoword was separated by 47 intervening sentences (approximately 25 minutes). In this condition, each sentence was presented for 30 seconds. After the time limit that included the time for reading and entering a meaning guess (i.e., 90 seconds for the massed condition and 30 seconds for the spaced condition), the correct meaning of the pseudoword was presented as feedback in the form of L1 (Japanese) translation equivalent and L2 (English) synonym. Each feedback episode was displayed for 30 seconds in the massed condition and 10 seconds in the spaced condition. The spacing regime for the massed and spaced learning conditions is detailed in Appendix A.

In incidental contextual vocabulary learning, learners do not always have a chance to see the correct meaning after making a contextual meaning inference. In this study, however, the correct meaning of the pseudoword was presented after inference attempts for two reasons. First, the presentation of the correct meaning as feedback emulates an authentic learning scenario in which L2 learners verify the meaning of unknown words by consulting dictionaries, asking their instructors or peers (Schmitt, 1997), or using glossaries. Second, research suggests that verifying contextually inferred meanings of unfamiliar words using dictionaries or glossaries increases vocabulary learning (Elgort, Beliaeva, & Boers, 2020; Laufer & Hulstijn, 2001; Yanagisawa, Webb, & Uchihara, in press). Examining whether

Accepted version

spacing further facilitates vocabulary learning from contextual meaning inferences followed by feedback is valuable because it may help researchers identify the optimal way to learn vocabulary from context.

Note that time-on-task was the same in the massed and spaced conditions; the only difference was in how the time was distributed. In both conditions, the participants spent 90 seconds guessing the meaning and 30 seconds viewing the feedback for each pseudoword (massed: 90-second guessing trial + 30-second feedback; spaced: 30-second guessing trial × 3 + 10-second feedback × 3). The whole treatment took 96 minutes. Participants were allowed to take a short break after every 15 sentences (an equivalent of 10 minutes). Unlike Koval (2019), the participants were not explicitly instructed to learn the pseudowords, nor were they forewarned about the upcoming posttests. As a result, the treatment in this study involved incidental (Hulstijn, 2001) or incidentally-oriented (Barcroft, 2015b) vocabulary learning.[2] After the treatment, participants completed a background questionnaire and performed 50 additions and subtractions (e.g., 53 + 43 = ?) as a filler task.

After the filler task, participants completed the lexical decision procedure with semantic priming (see Posttest Design and Dependent Measures). The experimental procedure was coded and delivered using E-Prime® software (Psychological Software Tools, Inc., Pittsburgh, PA). The task was conducted using a Fujitsu personal computer (Intel® Core™ i7 quad core CPU) with an LCD monitor (screen area: 1,280 × 1,024 pixels; refresh rate: 60 Hz) and the Chronos® response box. Participants were seated in front of the computer. They were instructed to indicate whether the string of letters presented on the screen (e.g., *predict, dapson, wreem*) was an English word or not, as quickly and accurately as possible, by pressing *yes* or *no* button on the response box. The prime – target pairs were presented one stimulus at a time (using a list-wise stimulus presentation); so, the participants made lexical decisions on each prime and target separately. Each stimulus was displayed until the participant registered a lexical decision, with an inter-trial interval of 200 milliseconds (e.g., Elgort, 2011; McRae & Boisvert, 1998). This list-wise presentation combined with a very low proportion of related prime – target pairs is considered preferable because it dramatically reduces the likelihood of semantic priming being influenced by deliberate task-

Accepted version

related processing or metacognitive strategies. The order of prime – target pair sequences was pseudorandomized in each list. Following the semantic priming task, an immediate meaning recall posttest was given. The test was administered using custom software created with Microsoft Visual Basic; pseudowords were presented in a different randomized order for each participant.

Session 2 (the delayed posttests) was conducted 2 days later. In Session 2, the semantic priming task was administered first, followed by the meaning recall posttest, followed by the meaning-form matching posttest. With the exception of the randomized item order, the delayed meaning recall posttest was identical to the immediate posttest. After the meaning-form matching posttest, a Japanese version of Operation Span (O-Span) task (Kobayashi and Okubo, 2014) was given to measure participants' working memory capacity. Following the O-Span task, the participants completed the background questionnaire and were asked to provide their opinions about the study.

Scoring and Data Analysis

To maintain consistency in scoring, the learners' meaning inferences during the learning phase were first scored by a computer program based on an answer key compiled by the first author and a research assistant (second-year MA student in applied linguistics), both native speakers of Japanese. The meaning inferences were categorized by the computer software into the following four categories: (a) correct responses, (b) blank responses, (c) cross-association errors (producing correct responses for other pseudowords), and (d) other responses. The responses that were classified as (d) other (14% of all responses) were independently scored by the first author and research assistant. The inter-rater agreement was 99.8% for the responses in category (d). The responses on the meaning recall posttest were scored using the same procedure; 3% of responses were categorized as (d) other and the inter-rater agreement was 100%. For the purposes of the data analysis, inference accuracy and meaning recall scores were treated as binary variables (i.e., correct or incorrect). Accuracy and RT of lexical decisions was recorded and scored by the E-Prime® software. An accuracy score of 1 was assigned for correct responses and 0 for incorrect responses. RT was measured from the stimulus onset until button press that registered a lexical decision.

Accepted version

Statistical analyses were conducted using mixed-effects regressions to examine the effect of the item presentation schedule (Schedule: Massed / Spaced) and the meaning inference accuracy (Guess.ACC: Correct / Incorrect) on the development of explicit and tacit word knowledge. Mixed logit models were used to analyze the following binary data: the meaning inferences accuracy, the accuracy of responses on the meaning recall and meaning-form matching posttests, and the accuracy of lexical decisions in the semantic priming task. Linear mixed-effects models were fitted to the response time data in the semantic priming task. We inverse-transformed response times (i.e., –1,000/response time) because the distribution of the nontransformed response times did not fit the assumption of normal distribution. We performed minimum a priori outlier removal (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, Davidson and Bates, 2008; Brysbaert and Stevens, 2018).

The data analysis was conducted using the lme4 package (version 1.1-17; Bates et al., 2015) in R (version 3.4.4; R Core Team, 2018). 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 et al., 2008; Cunnings, 2012). The primary interest predictors (Schedule and Guess.ACC), and an interaction between them were tested first (in that order), in the analyses of explicit knowledge. In the semantic priming analysis of tacit knowledge of meaning, Experimental Condition (related/unrelated) was a primary interest predictor tested first, followed by Schedule and Guess.ACC, and their two-way interactions. The following secondary interest variables were initially tested in all models: participants' vocabulary size in English (VST), Ospan, Age, and Theme (building/cooking), in that order. Session (immediate/delayed) was also included when the posttest was repeated, and its interaction with other predictors was tested. Because prior studies have shown that the L2 lexical processing may be less automatic, we included the number of letters as another secondary interest predictor in the tacit knowledge analyses. Finally, because RT and accuracy of

Accepted version

responses to primes may affect responses to the targets that follow them, the effect of these two variables, and their interaction with priming, were also tested in the RT analysis of lexical decisions to the targets.

The resulting models contained only variables that reached significance as predictors (i.e., their regression weights were significantly different from zero), improved the model fit, or were involved in interactions; all other predictors were excluded from further analysis. The final models contained random slopes supported by the data (i.e., parsimonious mixed models based on Matuschek et al., 2017). To control for the Type I error rate, the function *glht* in the R package multcomp (Hothorn, Bretz and Westfall, 2008) was used to obtain multiplicity-adjusted *p*-values.

Results

Analysis of Inference Accuracy During the Learning Phase

We compared inference accuracy for the massed schedule (single attempt) with that for each of the individual inference attempts (1, 2 and 3) in the spaced schedule (Figure 1, Table 1). Inference accuracy was the highest on the third attempt of the spaced treatment and it was significantly better than inference accuracy in the massed treatment ($z = 2.74$, multiplicity adjusted $p = .017$). However, inference accuracy in the massed treatment was significantly better than that on the first ($z = –15.65$, multiplicity adjusted $p < .001$) and second ($z = –6.29$, multiplicity adjusted $p < .001$) attempt of the spaced treatment. The analysis also showed that the participants with larger L2 vocabularies were better able to infer the meanings of the pseudowords from context ($d = 1.49$).

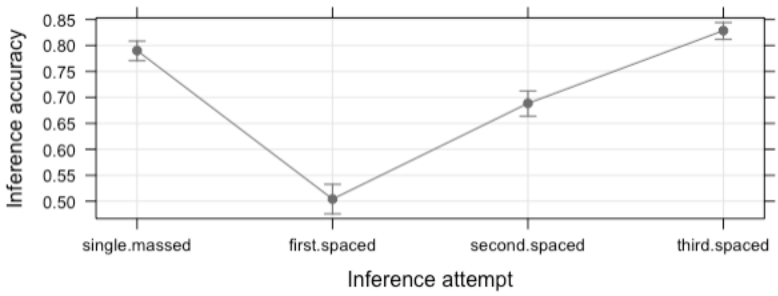

Figure 1. Effect plot for the inference accuracy by inference attempt (fit and 95% CIs).

Accepted version

Table 1. Accuracy of meaning inferences during the treatment phase (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d*
Intercept[a]	1.33	0.13	10.25	< .001	
Inference order = first, spaced	–1.31	0.08	–15.65	< .001	–0.72
Inference order = second, spaced	–0.53	0.08	–6.29	< .001	–0.29
Inference order = third, spaced	0.25	0.09	2.74	.006	0.14
Vocabulary size test (VST.lg.c)[b]	2.70	0.49	5.53	< .001	1.49

Note. [a]Intercept levels: Inference order = single, massed. [b]Vocabulary size test score, log-transformed, centered.

Analysis of Responses in the Meaning Recall Posttest

In the analysis of meaning recall (Figure 2, Table 2) there was a significant interaction between Schedule and Inference accuracy: in the spaced treatment, meaning recall was better when the final inference during treatment was correct; inference accuracy did not affect accuracy of meaning recall in the massed treatment.

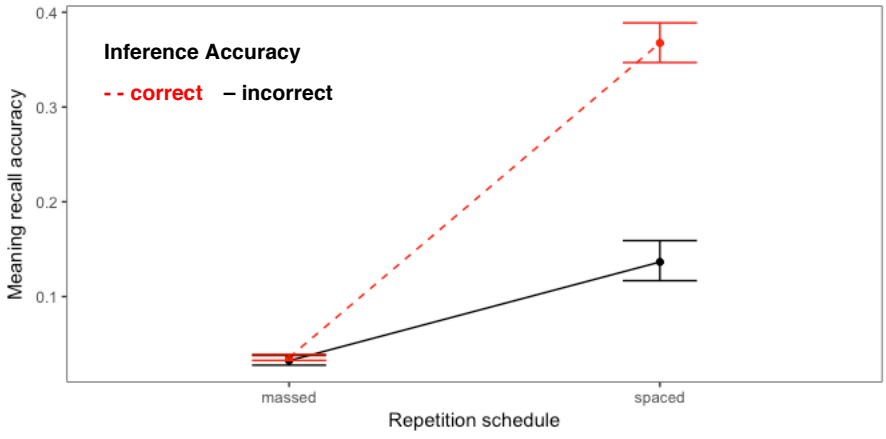

Figure 2. Effect plot for the interaction between Schedule and Inference accuracy in the analysis of meaning recall (fit and 95% CIs).

Importantly, there was also a large main effect of Schedule (d = 0.86), with around 27% advantage for the spaced treatment over the massed treatment (based on the model fit).

Accepted version

Meaning recall was also more accurate in the immediate than the delayed posttest, and for the cooking than the building theme (Table 2).

Table 2. Accuracy of meaning recall (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d*
Intercept[a]	–4.09	0.32	–12.94	< .001	–2.26
Schedule=spaced	1.55	0.27	5.81	< .001	0.86
Inference accuracy=1	0.10	0.23	0.45	.652	0.06
Session=immediate	0.72	0.08	9.31	< .001	0.40
Theme=cooking	0.67	0.28	2.37	.018	0.37
Schedule=spaced:Infer.accuracy=1	1.20	0.24	4.91	< .001	0.66

Note. [a]Intercept levels: Schedule = massed, Inference accuracy = 0 (incorrect), Session = delayed, Theme = building.

Analysis of Responses in the Meaning-form Matching Posttest

In the analysis of meaning-form matching there was a significant interaction between Schedule and Inference accuracy (Table 3, Figure 3): in the spaced treatment, accuracy of meaning-form matching was better when the final inference was correct; inference accuracy did not affect accuracy of meaning-form matching in the massed treatment.

Accepted version

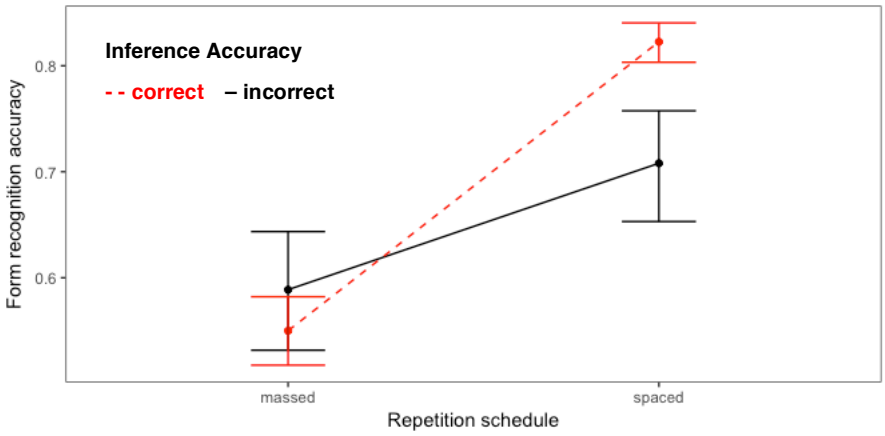

Figure 3. Effect plot for the interaction between Schedule and Inference accuracy in the analysis of meaning-form matching (fit and 95% CIs).

Importantly, there was also a main effect of Schedule ($z = 2.79$, $p = .005$), with around 24% advantage for the spaced over massed treatment (based on the model fit). Meaning recall was more accurate in the immediate than the delayed posttest, and for the cooking than building theme (Table 3).

Table 3. Analysis of the accuracy of meaning-form matching (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d*
Intercept[a]	0.57	0.21	2.73	.006	
Schedule=spaced	0.53	0.19	2.79	.005	0.29
Inference accuracy=1	–0.16	0.15	–1.07	.285	–0.09
Theme=home	–0.42	0.20	–2.13	.034	–0.23
Schedule=spaced: Infer.accuracy =1	0.81	0.21	3.83	< .001	0.45

Note. [a]Intercept levels: Schedule = massed, Inference accuracy = 0 (incorrect), Session = delayed, Theme = building.

Analysis of Responses in the Semantic Priming Task

We first conducted the analyses of the accuracy and response times data of lexical decisions on the word targets. In these analyses, the experimental condition (i.e.,

Accepted version

related/unrelated) was treated as a primary interest predictor. We predicted faster responses to the word targets in the related condition compared to the unrelated condition (i.e., priming) for the pseudowords that had been integrated into the lexical semantic networks of the learners. The second primary predictor was Schedule. If the spacing schedule differentially affected the development of semantic knowledge of the pseudowords, we would expect to see an interaction between the experiential condition (Cond) and Schedule. In addition to the analysis of the lexical decisions to the target words, we also analyzed lexical decisions to the primes (i.e., the pseudowords and real words).

Semantic priming: Accuracy analysis

The analysis showed significant negative semantic priming ($z = 2.79$, $p = .003$), with responses in the unrelated condition being about 2% more accurate than on semantically related trials, but there was no significant interaction between priming and Schedule (Appendix B). There was no effect of inference accuracy in this analysis.

Semantic priming: Response time analysis

Incorrect responses to targets (and their corresponding primes) were removed prior to the data analysis (21% of the data points). The final model included three significant two-way interactions: (a) between the experimental condition and accuracy of lexical decisions to the prime, (b) between the experimental condition and speed of lexical decisions to the prime, and (c) between accuracy of lexical decisions to the prime and the posttest session (immediate/delayed) (Table 4). Importantly, there was also a main effect of the experimental condition: responses in the related condition were faster than in the unrelated condition; that is, we observed the semantic priming effect for the pseudowords. However, there was no effect of Schedule, nor was there an interaction between Schedule and the experimental condition.

Accepted version

Table 4. Semantic priming, response time analysis (fixed effects)

Parameter	*b*	*SE*	*df*	*t*	*p*	*d*
Intercept[a]	–1.25	0.04	105.52	–34.57	< .001	3.25
Cond=unrelated	0.16	0.02	994.69	6.81	< .001	0.41
Prime RT (inv.Prime.RT.c[b])	0.07	0.02	145.83	3.76	< .001	0.18
Prime accuracy (Prime.ACC=1)	–0.02	0.01	4403.99	–1.44	.151	0.06
Session=immediate	0.06	0.02	135.49	3.96	< .001	0.17
Target length (centered)	0.08	0.01	46.96	5.67	< .001	0.20
Cond= unrelated:invPrime.RT.c	0.07	0.02	1037.73	3.27	.001	0.18
Cond=unrelated:Prime.ACC=1	–0.11	0.02	4596.23	–4.55	< .001	0.29
Prime.ACC=1:Session=immediate	–0.06	0.02	4674.36	–4.21	< .001	0.17

Note. [a]Intercept levels: Condition = related, Prime accuracy = 0, Session = delayed.
[b]Response times to primes, inverse transformed and centered.

Lexical decisions to primes: Response accuracy analysis

There was a significant two-way interaction in the analysis of response accuracy between Item type and Schedule; as expected, responses to the known words were not affected by the spacing schedule, but responses to the pseudowords were (Figure 4, Table 5). The lexical decisions were more accurate when the pseudowords had been encountered in the spaced than in the massed treatment.

Accepted version

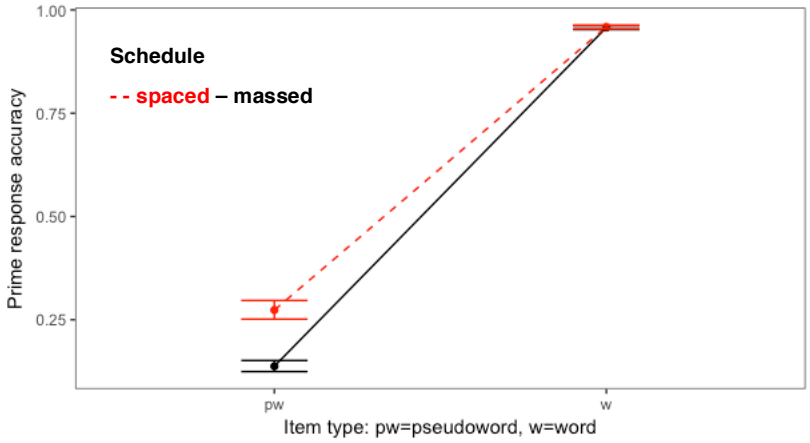

Figure 4. Effect plot for the interaction between Schedule and Item type in the analysis of accuracy of lexical decisions to primes (fit and 95% CIs).

Table 5. Lexical decisions to primes, response accuracy (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d*
Intercept[a]	–2.04	0.19	–11.00	< .001	–1.13
Item type=word	4.95	0.23	21.08	< .001	2.73
Schedule=spaced	0.86	0.09	9.19	< .001	0.47
Session=immediate	0.41	0.08	5.32	< .001	0.23
Item type=word:Schedule=spaced	–0.82	0.17	–4.79	< .001	–0.45

Note. [a]Intercept levels: Item type = pseudoword, Schedule = massed, Session = delayed.

Lexical decisions to primes: Response time analysis

Incorrect responses to primes were not removed from the analysis; instead, prime accuracy was included in the model. There was no effect of Schedule in the analysis of response times to primes (for the results, see Appendix C). The only result of note in this analysis was a two-way interaction between Item type and Session ($t = –8.48$, $p < .001$, $d = 0.25$); response times to the word primes were similar in the immediate and delayed sessions but responses to the pseudowords were faster in the delayed session.

Accepted version

Summary of Findings

We have compared the effect of two repetition schedules on contextual word learning. In the massed schedule, participants first read all three sentences with the target pseudoword presented on the same screen (no lag), then inferred its meaning and then reviewed the correct meaning. In the spaced condition, the participants read each of the three sentences separately with a lag of about 25 minutes; they inferred the meaning of the pseudoword after each encounter and reviewed the correct meaning after each inference attempt. We found that the participants were better able to infer the meanings of the pseudowords at the end of the spaced treatment than in the massed treatment. However, the first and the second meaning inference attempts in the spaced treatment were less accurate than in the massed treatment. This suggests that spaced contextual word learning is incremental and is facilitated through an inference-feedback loop that involves making contextual inferences, retrieving knowledge from previous learning episodes, and processing feedback.

Both explicit knowledge posttests (meaning-form matching and meaning recall) exhibited the spacing effect: better learning and retention was observed in the spaced than in the massed treatment, especially when the last inference in the spaced treatment was correct. Lexical decisions to the pseudoword primes were also more accurate when they were encountered in the spaced treatment, pointing to more precise lexical representations. We did not observe the spacing effect on the tacit knowledge, operationalized as semantic priming.

Discussion

In the present study, spaced distribution led to significantly higher scores than massed distribution on explicit knowledge posttests (meaning-form matching and meaning recall). The findings suggest that the spacing effect can be observed not only in decontextualized but also contextual vocabulary learning. The advantage of the spaced distribution over massed distribution might also be due in part to increased opportunities for retrieval and incremental access to feedback after each contextual inference. In the spaced condition, three sentences involving a given pseudoword were presented one by one, after approximately 25 minutes each. Each contextual encounter was accompanied by an explicit

Accepted version

inference attempt and feedback. The second and third presentations also allowed learners to retrieve information about the pseudoword learned in previous encounters. In the massed condition, in contrast, three sentences for a given pseudoword were presented simultaneously and constituted only one learning episode with one explicit inference attempt followed by feedback. Thus, the massed condition provided considerably reduced retrieval opportunities rather than a full inference-retrieval-feedback learning cycle. The superiority of spaced distribution over massed distribution in this study is consistent with earlier research demonstrating positive effects of retrieval and feedback on L2 vocabulary learning (e.g., Barcroft, 2007; Karpicke and Roediger, 2008; van den Broek et al., 2018).

Although spaced distribution was significantly more effective than massed distribution on posttests measuring explicit knowledge (meaning-form matching and meaning recall), the present study demonstrated no significant advantage of spaced distribution for the acquisition of tacit knowledge. The differential effect of the repetition schedule on type of knowledge in this study confirms that the acquisition of explicit and tacit knowledge may be affected by different factors (Bird, 2010; Elgort et al., 2018; Elgort and Warren, 2014; Suzuki and DeKeyser, 2017).

Another explanation for the lack of spacing effect on the acquisition of tacit knowledge is that the simultaneous presentation of three sentences in the massed condition may have encouraged (re)activation of the core senses of the pseudoword presented in three different informative contexts on the same screen, strengthening semantic connections of the novel vocabulary item with known L2 words, thus offsetting the spacing effect. In other words, although the spaced condition resulted in larger gains in the knowledge of explicit form-meaning mapping (measured by the posttests that afford explicit retrieval approaches), no spacing effect was observed in the semantic priming task relying on the online activation of semantic connections. This result is in line with the instance-based framework of word learning (Bolger et al., 2008), which suggests that multiple encounters with a word in supportive contexts result in the establishment of its semantic representation, as the word's core semantic features become abstracted from specific contexts.

Accepted version

Notably, the advantage of spacing was not observed in initial stages of learning. The first and second meaning inference attempts in the spaced treatment were less accurate than in the massed treatment, with only the third, final meaning inference being more accurate in the spaced than massed treatment. This corroborates the finding from the contextual word learning literature that multiple encounters are needed to acquire a word from reading (Elgort et al., 2018; Elgort and Warren, 2014; Pellicer-Sánchez, 2016; Pellicer-Sánchez and Schmitt, 2010; Webb, 2007). Our study shows contextual word learning both as a dynamic process and an outcome, for the two repetition schedules. The initial superior accuracy of meaning inferences in the massed condition may have been due to a more successful inductive learning from multiple simultaneously-presented examples of item use in context, providing more contextual clues than a single encounter with the word. This advantage diminished, however, by the second occurrence of the item in the spaced schedule, and was reversed by the third, as a result of the repeated distributed inference and retrieval attempts followed by feedback. At posttest, the advantage of the spaced treatment for the development of explicit knowledge of meaning and form-meaning mapping was clear.

We also observed a significant interaction between inference accuracy and schedule on the meaning recall and meaning-form matching posttests. In the spaced condition, the correct response on the last (third) inference attempt during the learning phase was associated with the correct response on the posttest, whereas there was no significant relationship between the inference accuracy and posttest performance in the massed condition. One explanation is the differential effects of retrieval and inference success on vocabulary learning. Memory research suggests that successful retrieval leads to better retention than unsuccessful retrieval because successfully recalling information strengthens a retrieval route, facilitating subsequent retrieval (e.g., Baddeley, 1997). However, the provision of feedback counteracts possible negative effects of initial incorrect retrieval (Carpenter et al., 2012; Elgort et al., 2020). Recall that the correct response on the inference attempt in the massed condition was due purely to inference success whereas the correct response on the third attempt in the spaced condition may have been caused by a combination of inference and retrieval success and the effect of feedback. This may explain why the learners' posttest

Accepted version

performance in the spaced, but not massed, condition was mediated by the accuracy of the final inference in the learning phase.

Another plausible explanation is that the inference accuracy for the third retrieval attempt in the spaced condition reflects the learning burden of the pseudowords whereas the single inference attempt in the massed condition does not. Because in the spaced condition the learners reviewed the correct meaning of the pseudoword after submitting a contextual meaning inference twice before making the final, third inference, an inference error on the third attempt suggests that the pseudoword was perhaps more difficult to learn, i.e., was associated with a greater learning burden. In the massed condition, learners were not exposed to the correct meanings of the pseudowords prior to the single inference attempt; therefore, inference accuracy in the massed treatment indexed the guessability (ease of meaning inferences) rather than the learnability of the pseudowords. This may explain why the accuracy of explicit pseudoword knowledge on the posttests was associated with the inference accuracy in the spaced but not massed learning condition.

Theoretical Account of the Findings

A number of theoretical frameworks have been proposed to explain the spacing effect, such as the encoding variability account (e.g., Maddox, 2016), deficient processing account (e.g., Koval, 2019), and transfer appropriate processing account (e.g., Russo & Mammarella, 2002). According to the encoding variability account, in a massed schedule, information tends to be encoded in a stable, fixed context, whereas in a spaced schedule, it is encoded in more physically, mentally, or temporally diverse contexts. The encoding variability account suggests that more diverse encoding processes in a spaced schedule facilitate later recall. The deficient processing account states that information presented in a spaced schedule receives more attention or processing than information presented in a massed schedule, which results in superior retention. Transfer appropriate processing theory suggests that memory performance is enhanced if there is a close match between the context of encoding and that of testing (Morris, Bransford, & Franks, 1977). When applied to the spacing effect, this theory predicts that spaced learning, where information is presented over

Accepted version

a longer period of time, results in the kind of knowledge that can be accessible for a long time, whereas massed learning may facilitate only short-term memory.

The findings of the present study cannot be fully accounted for by either the encoding variability or deficient processing account because neither of the two frameworks predicts the advantage of spacing over massing for the acquisition of explicit, but not tacit knowledge. The results of this study may be better explained by the transfer appropriate processing account. In the spaced (but not in the massed) condition, participants had multiple, distributed opportunities to retrieve the meaning of the previously encountered target items. Note that explicit measures used in this study (i.e., meaning recall and meaning-form matching posttests) also required learners to retrieve the meaning of the target items. However, neither spaced nor massed sentence reading treatment required the kind of processing tapped into by the semantic priming task, that is, implicit processing of the target items and words related to them in their meaning (rather than words that co-occurred with them in the text). The transfer appropriate processing account, therefore, predicts better performance for the spaced than massed condition on the posttests of explicit knowledge of meaning, but not on the measure of tacit knowledge of meaning, as is the case in the present study. However, because the transfer appropriate account of the spacing effect has not received as much attention from researchers as the encoding variability or deficient processing account, more empirical studies need to be conducted to test the validity and explanatory power of this account.

Practical Implications of the Findings

A key finding of our study is that, in L2 vocabulary acquisition from reading, the development of explicit knowledge of form and meaning (and their mapping) is facilitated when new words occur with a lag rather than when they are clumped together within the text, particularly when learners are able to verify their contextual inferences, e.g., using a dictionary or glossary. This insight is useful for content developers (both publishers and teachers) because it can help them create or select more effective materials for L2 learners to build their vocabulary from reading. In addition, we found that tacit semantic knowledge develops from encountering new words in supportive context, irrespective of whether the

Accepted version

new words occur in a massed or spaced manner in the text, which emphasizes the importance of reading for vocabulary development.

Although some researchers argue that pure massing rarely takes place in a classroom setting and is not very educationally relevant (e.g., Rogers and Cheung, in press), we maintain that massing does in fact occur in L2 teaching. For example, in concordance-based learning activities recommended by a number of L2 vocabulary scholars (e.g., Nation, 2013; Schmitt, 2000), multiple instances of a new word are presented in context on the same page. Thus, learning words from a concordance output is essentially massed learning. Although concordance-based activities can help learners notice patterns of language use (for example, how words co-occur) or generate explicit meaning inferences, our findings suggest that concordance outputs will be less effective in learning form-meaning connections. Spaced encounters with new words in supportive contexts are likely to result in better learning outcomes, at least for the learning of form-meaning connections.

Massing is also a common way of presenting items in contextual L2 vocabulary learning research. Based on the findings of the present study, this may not be an optimal learning choice. We recommend that future vocabulary learning studies deliberately consider the effect of spacing at the research design stage.

Concluding Remarks

This study examined the effects of massed and spaced distributions on the acquisition of explicit and tacit vocabulary knowledge from reading. Our study design emulated an authentic learning scenario in which L2 learners verify their contextual meaning inferences of unknown words by consulting dictionaries, asking their instructors or peers (Schmitt, 1997), or using glossaries. The results showed that, when contextual meaning inferencing is required and is followed by feedback (i.e., L2 synonyms and L1 translation equivalents), spaced learning is more effective for the acquisition of explicit knowledge of form-meaning mapping than massed learning. However, the spacing advantage may not hold without the provision of feedback after each inference attempt. A combination of spaced retrieval and feedback is known to facilitate L2 vocabulary learning (e.g., Barcroft, 2007; Karpicke and Roediger, 2008; van den Broek et al., 2018); therefore, our design may have

Accepted version

worked in favor of the spaced treatment, which involved opportunities for retrieval. In future research, the effects of massing and spacing on contextual vocabulary learning should also be examined without access to feedback in order to verify whether vocabulary learning from reading without access to reference materials is differentially affected by spacing.

Another useful follow-up from this study would be an investigation of how spacing affects tacit knowledge in decontextualized paired-associate learning. In our study, spacing did not affect the development of tacit semantic knowledge. We hypothesized that, by presenting multiple sentences simultaneously, massing may have allowed learners to establish a more robust set of semantic features that they could access in the online semantic priming posttest. In decontextualized learning, however, massing does not offer the same advantage; thus spacing should be more effective than massing for the acquisition of tacit semantic knowledge. Considering that existing decontextualized studies of the spacing effect only measured explicit knowledge (e.g., Karpicke and Bauernschmidt, 2011; Nakata, 2015; Seibert, 1932), further research comparing the effects of massing and spacing on the acquisition of tacit semantic knowledge in decontextualized learning is warranted.

Compared with the lag effect, which is sometimes referred to as "nebulous" (Rogers, 2017, p. 907), the spacing effect is considered very robust in associative learning, and L2 vocabulary researchers have recommended a spaced learning approach (e.g., Barcroft, 2015b; Ellis, 1995; Hulstijn, 2001; Nation, 2013; Schmitt, 2000).[3] Our study makes initial steps in extending the spacing effect to contextual vocabulary learning, as far as the acquisition of explicit knowledge is concerned. On the other hand, massing appears to be just as effective as spacing for the acquisition of tacit knowledge. Methodologically, our findings suggest that when comparing the effects of massing and spacing, it is important to measure both explicit and tacit knowledge because tacit knowledge of meaning underpins fluent word-to-text integration. Further research examining the effects of spacing on the acquisition of tacit knowledge is warranted because rich and flexible tacit semantic knowledge is essential for fluent language processing.

Accepted version

Notes

[1] In some studies, the term *massed learning* is used to refer to relatively short spacing, whereas *spaced learning* refers to relatively long spacing. In this study, massed learning is used to refer to a practice schedule that does not involve any spacing (Cepeda et al., 2006).

[2] Hulstijn (2001) defines incidental vocabulary learning as an activity where participants are not explicitly instructed to learn target vocabulary items and are not aware of upcoming vocabulary posttests. Barcroft (2015b), however, argues that the lack of forewarning of vocabulary posttests does not necessarily guarantee that learners do not attempt to learn target vocabulary items intentionally and considers it more appropriate to use the term *incidentally-oriented vocabulary learning* instead of incidental vocabulary learning.

[3] For non-L2 vocabulary research failing to show the advantage of spacing over massing (i.e., *Peterson paradox*), see a meta-analysis conducted by Cepeda and colleagues (2006).

Acknowledgements

An earlier version of this paper was presented at Vocab@Leuven Conference and Japan Second Language Acquisition Research Forum Second Meeting. This research was supported in part by JSPS Grant-in-Aid for Research (#16H05943) awarded to the first author. We are very grateful to Shotaro Ueno, Saki Nagamine, and Hiroko Kashiba for their assistance with data collection. We are also very grateful to two anonymous reviewers, Anna Siyanova, and the SLR Editors for their invaluable feedback on earlier versions of the manuscript.

References

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

Baayen RH, Davidson DJ and Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Baddeley AD (1997) *Human memory: Theory and practice*. Revised ed. East Sussex, UK, Psychology Press.

Accepted version

Bahrick HP and Phelps E (1987) Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 344–349.

Barcroft J (2007) Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57, 35–56.

Barcroft J (2015a) Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, 48, 236–249.

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

Bates D, Mächler M, Bolker B, et al. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Bird S (2010) Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, 31, 635–650.

Bolger DJ, Balass M, Landen E, et al. (2008) Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45, 122–159.

Brysbaert M and Stevens M (2018) Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1, 9.

Carpenter SK, Sachs RE, Martin B, et al. (2012) Learning new vocabulary in German: The effects of inferring word meanings, type of feedback, and time of test. *Psychonomic Bulletin and Review*, 19, 81–86.

Cepeda NJ, Pashler H, Vul E, et al. (2006) Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354–380.

Cunnings I (2012) An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28, 369–382.

Ebbinghaus H (1885) Memory: A contribution to experimental psychology. New York, NY, Genesis Publishing.

Elgort I (2011) Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61, 367–413.

Accepted version

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.

Elgort I, Brysbaert M, Stevens M, et al. (2018) Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40, 341–366.

Elgort, I., Beliaeva, N., & Boers, F. (2020). Contextual word learning in the first and second language: Definition placement and inference error effects on declarative and nondeclarative knowledge. *Studies in Second Language Acquisition, 42*, 7-32. doi:10.1017/S0272263119000561

Elgort I and 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.

Ellis NC (1995) The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, 8, 103–128.

Hothorn T, Bretz F and Westfall P (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, 50, 346–363.

Hulstijn JH (2001) Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In: Robinson P (ed.), *Cognition and second language instruction*, Cambridge, UK, Cambridge University Press, pp. 258–286.

Karpicke JD and Bauernschmidt A (2011) Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1250–1257.

Karpicke JD and Roediger HL (2008) The critical importance of retrieval for learning. *Science*, 319, 966–968.

Kobayashi A and Okubo M (2014) 日本語版オペレーションスパンテストによるワーキングメモリの測定(Assessment of working memory capacity with a Japanese version of the Operation Span Test). *The Japanese Journal of Psychology*, 85, 60–68.

Accepted version

Kornell N and Bjork RA (2008) Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, 19, 585–592.

Koval NG (2019) Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, 40, 1103–1139.

Kurtz KH and Hovland CI (1956) Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, 51, 239–243.

Laufer B (2003) Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, 59, 567–587.

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

Maddox GB (2016) Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, 28, 684–706.

Matuschek H, Kliegl R, Vasishth S, et al. (2017) Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.

McRae K and Boisvert S (1998) Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 558–572.

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

Morris CD, Bransford JD and Franks JJ (1977) Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16, 519–533.

Nakata T (2015) Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37, 677–711.

Accepted version

Nakata T and Suzuki Y (2019) Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*. 41, 287–311.

Nation ISP (2013) *Learning vocabulary in another language*. 2nd ed. Cambridge, UK, Cambridge University Press.

Nation ISP and Beglar D (2007) A vocabulary size test. *The Language Teacher*, 31(7), 9–13.

Pellicer-Sánchez A (2016) Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, 38, 97–130.

Pellicer-Sánchez A and Schmitt N (2010) Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, 22, 31–55.

Rogers J (2017) The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, 38, 906–911.

Rogers J and Cheung A (in press) Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*.

Russo R and Mammarella N (2002) Spacing effects in recognition memory: When meaning matters. *European Journal of Cognitive Psychology*, 14, 49–59.

Schmitt N (1997) Vocabulary learning strategies. In: Schmitt N and McCarthy M (eds.), *Vocabulary: Description, acquisition and pedagogy*, Cambridge, UK, Cambridge University Press, pp. 199–227.

Schmitt N (2000) *Vocabulary in language teaching.* Cambridge, UK, Cambridge University Press.

Seibert LC (1932) *A series of experiments on the learning of French vocabulary*. Baltimore, MD, The Johns Hopkins Press.

Serrano R and Huang H-Y (2018) Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, 52, 971–994.

Suzuki Y and DeKeyser R (2017) Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, 21, 166–188.

Accepted version

Tinkham T (1997) The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research,* 13, 138–163.

van den Broek GSE, Takashima A, Segers E, et al. (2018) Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning*, 68, 546–585.

Webb S (2007) The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28, 46–65.

Webb S and Chang A (2015) Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, 19, 667–686.

Yanagisawa A, Webb S and Uchihara T (in press) How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*.

Zulkiply N and Burt JS (2013) The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory and Cognition*, 41, 16–27.

Accepted version

Appendix A. Sample item order during the treatment

The following figure gives a sample item order during the treatment. In the figure, "Massed 1" refers to the first pseudoword assigned to the massed condition, and "Spaced 1" refers to the first pseudoword assigned to the spaced condition.

Position	Item	Sentences
1	Massed 1	1, 2, 3
2	Spaced 1	1
3	Spaced 2	1
4	Spaced 3	1
5	Massed 2	1, 2, 3
6	Spaced 4	1
7	Spaced 5	1
8	Spaced 6	1
9	Massed 3	1, 2, 3
10	Spaced 7	1
11	Spaced 8	1
12	Spaced 9	1
13	Massed 4	1, 2, 3
14	Spaced 10	1
15	Spaced 11	1
16	Spaced 12	1
17	Massed 5	1, 2, 3
18	Spaced 13	1
19	Spaced 14	1
20	Spaced 15	1
21	Massed 6	1, 2, 3
22	Spaced 16	1
23	Spaced 17	1
24	Spaced 18	1
25	Massed 7	1, 2, 3
26	Spaced 19	1
27	Spaced 20	1
28	Spaced 21	1
29	Massed 8	1, 2, 3
30	Spaced 22	1
31	Spaced 23	1
32	Spaced 24	1
33	Massed 9	1, 2, 3
34	Spaced 1	2
35	Spaced 2	2
36	Spaced 3	2
37	Massed 10	1, 2, 3
38	Spaced 4	2
39	Spaced 5	2
40	Spaced 6	2
41	Massed 11	1, 2, 3
42	Spaced 7	2
43	Spaced 8	2
44	Spaced 9	2
45	Massed 12	1, 2, 3
46	Spaced 10	2
47	Spaced 11	2
48	Spaced 12	2
49	Massed 13	1, 2, 3
50	Spaced 13	2
51	Spaced 14	2
52	Spaced 15	2
53	Massed 14	1, 2, 3
54	Spaced 16	2
55	Spaced 17	2
56	Spaced 18	2
57	Massed 15	1, 2, 3
58	Spaced 19	2
59	Spaced 20	2
60	Spaced 21	2
61	Massed 16	1, 2, 3
62	Spaced 22	2
63	Spaced 23	2
64	Spaced 24	2
65	Massed 17	1, 2, 3
66	Spaced 1	3
67	Spaced 2	3
68	Spaced 3	3
69	Massed 18	1, 2, 3
70	Spaced 4	3
71	Spaced 5	3

Accepted version

72	Spaced 6	3
73	Massed 19	1, 2, 3
74	Spaced 7	3
75	Spaced 8	3
76	Spaced 9	3
77	Massed 20	1, 2, 3
78	Spaced 10	3
79	Spaced 11	3
80	Spaced 12	3

81	Massed 21	1, 2, 3
82	Spaced 13	3
83	Spaced 14	3
84	Spaced 15	3
85	Massed 22	1, 2, 3
86	Spaced 16	3
87	Spaced 17	3
88	Spaced 18	3
89	Massed 23	1, 2, 3

90	Spaced 19	3
91	Spaced 20	3
92	Spaced 21	3
93	Massed 24	1, 2, 3
94	Spaced 22	3
95	Spaced 23	3
96	Spaced 24	3

As shown above, the three sentences from the massed and spaced conditions alternated throughout the treatment. For instance, at the beginning of the treatment, three sentences for the first item in the massed condition (Massed 1) were presented for 90 seconds. After that, three sentences from the first three items in the spaced condition (Spaced 1–3) were presented one by one, for 30 seconds each. This was followed by three sentences for the second item in the massed condition (Massed 2) presented for 90 seconds.

The item order was randomized anew for each participant to minimize the order effect. For instance, for one participant, *shottle* (gravel 砂利) may be Massed 1 and *tenont* (pumpkin かぼちゃ) may be Massed 24, whereas for another participant, *dapson* (detergent 洗剤) may be Massed 1 and *brophy* (strainer ざる) may be Massed 24. The randomization was implemented in such a way that pseudowords from the two themes (building/household and cooking/food) were distributed roughly equally across the treatment. To control for the order effect, for half of the participants, the first three sentences were from the massed condition, and for the other half of the participants, the first three sentences were from the spaced condition.

Accepted version

Appendix B. Semantic priming, accuracy analysis (fixed effects)

Parameter	*b*	*SE*	*z*	*p*	*d*
Intercept[a]	1.73	0.26	6.73	< .001	0.95
Cond=unrelated	0.22	0.08	2.97	.003	0.12
Schedule=spaced	0.20	0.08	2.63	.009	0.11
Session=immediate	0.34	0.08	4.57	< .001	0.19
Vocabulary size (VST.lg.c[b])	4.01	1.06	3.77	< .001	2.21
Target length (Target.length.c[c])	–0.30	0.12	–2.64	.008	–0.17

Note. [a]Intercept levels: Condition = related, Schedule = massed, Session = delayed.

[b]Vocabulary size score, log-transformed, centered. [c]Target length in letters, centered.

Accepted version

Appendix C. Lexical decisions to primes, response times (fixed effects)

Parameter	*b*	*SE*	*df*	*t*	*p*	*d*
Intercept[a]	–1.29	0.04	99.16	–32.15	< .001	2.83
Item type=word	–0.03	0.03	118.79	–1.05	.296	0.08
Session=immediate	0.16	0.02	87.72	8.94	< .001	0.35
Prime accuracy (Prime.ACC)=1	0.14	0.03	86.43	4.95	< .001	0.30
Prime length (Prime.nol)=7	0.07	0.02	77.58	3.72	< .001	0.15
Item type=word:Session=imm.	–0.11	0.01	5886.22	–8.48	< .001	0.25
Item type=word:Prime.ACC=1	–0.26	0.04	96.02	–6.53	< .001	0.56

Note. [a]Intercept levels: Item type = pseudoword, Session = delayed, Prime accuracy = 0, Prime length (Prime.nol) = 6.

Accepted version