Learning Words via Reading: Contextual Diversity, Spacing and Retrieval Effects in Adults

Ascensión Pagán[1,2] and Kate Nation[1]

[1]University of Oxford

[2]University of Leicester

Author Note

Correspondence regarding this article should be addressed to Ascensión Pagán who is now at the Department of Neuroscience, Psychology and Behaviour, College of Life Sciences, University of Leicester, George Davies Centre, Lancaster Road, Leicester, LE1 7HA, UK or to Kate Nation, Department of Experimental Psychology, Anna Watts Building, Woodstock Road, Oxford, OX2 6GG, UK. E-mails:appc1@leicester.ac.uk; kate.nation@psy.ox.ac.uk.

Abstract

We examined whether variations in contextual diversity, spacing and retrieval practice influenced how well adults learned new words from reading experience. Eye movements were recorded as adults read novel words embedded in sentences. In the learning phase, unfamiliar words were presented either in the same sentence repeated four times (same context) or in four different sentences (diverse context). Spacing was manipulated by presenting the sentences under distributed or non-distributed practice. After learning, half of the participants were asked to retrieve the new words and half had an extra exposure to the new words. Although words experienced in diverse contexts were acquired more slowly during learning, they enjoyed a greater benefit of learning at immediate post-test. Distributed practice also slowed learning, but no benefit was observed at post-test. Although participants who had an extra exposure showed the greatest learning benefit overall, learning also benefited from retrieval opportunity, when words were experienced in diverse contexts. These findings demonstrate that variation in the content and structure of the learning environment impacts on word learning via reading.

Keywords: word learning, reading, eye movements, spacing, contextual diversity

1. Introduction

Text provides a powerful substrate from which to learn new words. From mid-childhood onwards, people learn more words via reading than through conversations (e.g., Nagy, Herman, & Anderson, 1985), yet relatively little is known about the factors that influence word learning via reading in adults. Natural text is complex and varied, and the meaning of a new word is rarely explicit. This contrasts with the simplicity of memory paradigms that use recall or recognition to probe verbal learning, and the explicitness of tasks such as referent matching. In this paper, we investigated whether variability during reading experience influenced how adults read new words embedded in text, and how well they learned them. We manipulated three different types of variability, each of which is associated with variation in learning or processing in other domains: contextual variability, temporal spacing and retrieval opportunity.

1.1.Contextual diversity

Evidence that contextual variation matters for verbal learning and verbal memory comes from a variety of sources (for review, Jones, Dye & Johns, 2018). For example, variability supports category learning in infants and adults (Goldenberg & Sandhofer, 2013; Smith & Handy, 2014). Infants show better retention of object labels if they experience object-label mappings against a changing background across trials, rather than one that remains constant across trials (Twomey & Westermann, 2017). In learning paradigms with adults, variation in context is associated with better learning of grammatical class (Reddington, Chater, & Finch, 1998) and referent mappings (Smith & Yu, 2008). These findings fit with classic effects in the verbal memory literature, where variation at study leads to superior recall in novel environments at test, presumably due to the formation of representations that are less context-dependent (e.g., Godden & Baddeley, 1975).

Turning to reading itself, Adelman, Brown & Quesada (2006) found an item-level relationship between how quickly a word is named or responded to in lexical decision and its contextual diversity. They indexed contextual diversity using document count –the number of unique documents a word appeared in across a large corpus of language and found document count to be a better predictor of lexical processing than frequency. Similarly, high document count words are fixated for less time in sentence reading than frequency-matched words lower in document count (Plummer, Perea, & Rayner, 2014). These findings suggest that the number of different contexts a word is experienced in influenced lexical processing, beyond the number of times it has appeared. One explanation for this is the principle of likely need. This refers to the idea that as contextually diverse words are more likely to be needed in new contexts, they become more accessible in semantic memory (Adelman *et al.*, 2006; Anderson & Milson, 1989; Anderson & Schooler, 1991).

1.2.Spacing

Repeated items are more likely to be recalled if repetitions at study are separated in time compared to when the repetitions occur one after the other (for review, see Janiszewski, Noel, & Sawyer, 2003). Similar effects hold for word learning (e.g., Bloom & Shuell, 1981; Cepeda, Pashler, Vul, Wixted, & Rohre, 2006; Challis, 1993; Greene, 1989; Hintzman & Block, 1973; Verkoeijen, Rikers, & Schmidt, 2004). A number of factors influence the size of the spacing effect, including whether learning is tested immediately after the study phase (acquisition) or 24 hours or more afterwards (retention), and the type of task (e.g., Cepeda *et al.*, 2006; Donovan & Radosevich, 1999). For example, with linguistic tasks, larger spacing effects are observed as the retention interval increases (e.g., Underwood & Ekstrand, 1967).

The specific cause of the spacing advantage is unclear, but it is possible to discern three types of account that could explain its benefit: activation, consolidation and contextual variation. Both the activation and consolidation accounts assume that the timing between presentations drives the spacing benefit (e.g., Landauer, 1969; Pavlik & Anderson, 2005, 2008). In contrast, the contextual variation account assumes that it is the nature of the context that appears with the item that is critical. On this view, the context that accompanies the item at each presentation is also encoded. Temporal spacing increases the likelihood that new contextual or episodic information is encoded with each repetition, thus strengthening the memory trace (e.g., Greene, 1989; Raaijmakers, 2003).

The effect of spacing on word learning via reading has not been investigated directly. Relevant, however, are findings reported by Betts, Gilbert, Cai, Okedara & Rodd (2018) who investigated how recent encounters with an ambiguous word in a specific context influences the availability of its meaning. They had participants listen to paragraphs which contained either one or three exposures to the subordinate meaning of an ambiguous word such as *bark*. This exposure led to word-meaning priming: the recently encountered meaning was more likely to be produced in a subsequent word association task, relative to baseline. The size of the priming effect was equivalent regardless of whether the recently encountered meaning had been heard three times or only once. This suggests that a single encounter is sufficient to retune lexical-semantic representations. In a follow-up experiment, participants again heard three exposures to a biased interpretation of an ambiguous word but this time in three different sentences, spaced into different blocks across the experiment. This spaced exposure resulted in more word-meaning priming, relative to both a single exposure and to three exposures massed together in the same block. This demonstrates that spacing leads to changes in lexical activation that influence comprehension some 20-30 minutes later. Plausibly, these experiences bring about longer-lasting changes in lexical-semantic knowledge, consistent with the view that word knowledge is dynamic and influenced by distributional factors beyond the number of times a word has been encountered.

1.3.Retrieval

It is well-accepted that testing and retrieval practice influences learning. For example, Hogan and Kintsch (1971) found that at test (48 hours after study), participants who had the opportunity to retrieve items showed better word learning than those who had an extra exposure (see also McDaniel & Mason, 1985; Roediger & Karpicke, 2006; Tulving, 1967; for review, see Roediger & Karpicke, 2006; Roediger & Butler, 2011).

How does retrieval influence memory? One explanation is that retrieving information generates a retrieval route which in turn makes successive retrievals more likely (e.g., Carpenter, 2009). Another possibility is that retrieval practice allows elaboration of the information to be learnt and this increase in processing effort leads to better learning in the longer term (e.g., Pyc & Rawson, 2009). Bjork & Bjork (1992) described two inversely related mechanisms to explain retrieval practice effects: storage strength and retrieval strength. Storage strength refers to how permanent the memory trace is while retrieval strength refers to its accessibility. Retrieval improves storage strength, and the effects of this are most beneficial when forgetting is most likely – repeated retrieval at a time when retrieval strength is relatively high (immediately after study, for example) will be less beneficial than at the point of forgetting.

1.4. Reading experience and word learning via reading

So far, we have provided a brief discussion of three different types of variability, namely contextual diversity, spacing and retrieval practice, and concluded that all three factors influence verbal learning and memory. With these findings as a backdrop, we now turn to consider the case of reading in more detail.

Accounts of word reading grounded in episodic memory provide a natural link to the memory literature. Reichle and Perfetti (2003) implemented an episodic model of visual word learning and word recognition based on MINERVA 2 (Hintzman, 1984). This model comprises memory traces along with operations to encode and retrieve them. Each memory trace represents a specific experience a reader has had with a word, capturing experience with its form and the context in which it was encountered. When words are encountered and recognized on subsequent occasions, new traces are added to those already encoded. The ease of retrieval within the model represents an item's availability which in turn influences recognition. Motivated by this model, Bolger, Balass, Landen and Perfetti (2008) devised a word learning experiment that manipulated contextual diversity during a learning phase. Participants read rare unfamiliar words embedded in meaningful sentences a number of times; the carrier sentences were either identical on each reading or varied in linguistic content each time; this manipulation is akin to variation in document count (e.g., Adelman *et al.*, 2006). At post-test, words experienced in varying contexts were better learned. Bolger *et al.* concluded that contextual variation during learning caused the novel words to become more context independent. This meant that they were more accessible and therefore more easily processed in new contexts, in line with predictions from Reichle and Perfetti's (2003) episodic model.

Moving away from the specifics of this model, Nation (2017) outlined how naturalistic reading experience across the lifespan provides immense exposure to words in different episodes and contexts; over time, these experiences sum to a rich and complex person-specific lexical history for words and the contexts in which they appear. The product of this experience is item-level variation in lexical quality, defined as the extent to which the mental representation of a word specifies its meaning and form (Perfetti, 2007). In turn, variations in lexical quality relate to variation in processing. Within this general framework, we can consider the factors that might influence word learning via reading experience. Our earlier discussion highlighted the effects of temporal spacing and retrieval opportunity on verbal learning and memory and Bolger *et al.*'s (2008) experiment showed the relevance of contextual diversity to this type of learning; it is plausible that such factors are also at play during word learning via reading. In natural language, however, it is difficult to isolate effects, not least because text is complex and reading experience is vast. For example, contextual diversity might reflect differences in temporal spacing over time, with words that occur in more unique documents becoming more context independent, perhaps as a result of differences in retrieval practice. In addition, item-level characteristics can be highly correlated. For example, Adelman *et al.* (2006) reported a correlation of r=.96 between contextual diversity as indexed by document count and word frequency.

Word learning paradigms offer a tightly controlled means to directly investigate how different types of variation in the environment lead to differences in learning and processing. Our goal in this experiment was to manipulate three features of reading experience: contextual diversity, temporal spacing and retrieval practice. These factors have not been examined collectively before and while contextual diversity has been a variable of interest in the reading literature (e.g., Bolger *et al.*, 2008; Johns, Dye & Jones, 2016; and in children, see Joseph & Nation, 2018), the effects of spacing and retrieval opportunity on new word learning via reading are unknown. To capture learning from reading experience, we adapted the word learning paradigm used by Joseph, Wonnacott, Forbes and Nation (2014) and further developed by Joseph and Nation (2018) in which eye movements are measured as people encounter new words in text. The basic premise is that as familiarity builds, fixation times generally become shorter. The eye movement record provides detailed information about how novel words are processed (e.g., Bai, Liang, Blythe, Zang, Yan, & Liversedge, 2013; Blythe, Liang, Zang, Wang, Yan, Bai, & Liversedge, 2012; Chaffin, Morris, & Seely, 2001; Englot, Brysbaert, Stevens, & Van Assche, 2017; Liang, Blythe, Zang, Bai, Yan, & Liversedge, 2015; Lowell & Morris, 2014); of interest is whether changes in fixation patterns reflect differences induced by the learning environment. More generally, the paradigm allows learning to be captured in naturalistic circumstances, as people read meaningful text silently and with no additional task requirements.

Our experiment comprised three phases, all completed in a single session: a pre-exposure phase which provided baseline data, a learning phase which manipulated features of the learning environment, and a post-exposure test phase that assessed learning. Eye movements were recorded throughout all three phases. For the pre-exposure phase, rare unfamiliar words were embedded in neutral sentences (i.e., sentences which provided no direct instruction as to their meaning). During the learning phase, the same rare words were presented but this time in contextual sentences which provided cues to their meanings. To manipulate contextual diversity, either the same word was read in the same sentence four times, or four different sentences containing the same word were each read once. Temporal spacing was manipulated such that words in the spaced condition were distributed across blocks with short 3-minute intervals between each presentation; in the massed condition, exposures occurred in successive blocks. To manipulate retrieval opportunity, one group of participants were asked to write down as many words they could remember from the learning phase while another group received an additional exposure to each target word, via an e-cancellation task. Finally, in the post-exposure phase, participants once again read the target words in different neutral sentences and then answered a comprehension question to assess knowledge of the new words directly.

Our predictions were as follows. Consistent with evidence from other learning and memory paradigms, we predicted that words experienced in diverse contexts would be better learned than words seen in repeated contexts. Thus, we anticipated diverse words would receive shorter fixation times than non-diverse words. Similarly, we anticipated that temporal spacing would also lead to better learning, resulting in shorter fixation durations, relative to words seen in the massed condition. Given evidence that active retrieval produces greater learning, we predicted that readers who had an opportunity to retrieve would show more learning, relative to participants who received an additional exposure to each target word. Our design also allowed us to examine interactions between diversity, spacing and retrieval.

2. Method

2.1.Participants

Sixty-eight adults aged 18-30 years (24 males, 44 females) took part in this experiment. They were recruited through the University of Oxford's online participant recruitment system. All were native speakers of English with normal or corrected to normal vision and no history of reading difficulty. Reading level was checked at the start of the experiment and all participants scored comfortably within age-expected levels according to the Test of Word Reading Efficiency, a standardised test of word reading fluency (Rashotte, Torgesen & Wagner (1999); sum of standard scores for words and nonwords, M = 218.4; SD = 17.7). Participants were unaware of the purpose of the experiment. They received £10 or course credits for participating. This experiment was approved by the University of Oxford Ethics Committee and all participants gave informed written consent.

2.2.Apparatus

The sentences were presented on a 15" Dell monitor, set at a refresh rate of 60 Hz with 1024 x 768 resolution, interfaced with a PC at a viewing distance of 60 cm. Sentences were presented in black, Courier New, size 12 font on a grey background; three characters subtended 1° of visual angle. Although reading was binocular, eye movements were recorded only from the right eye, using an SR Research EyeLink 1000 tracker (SR Research Ltd., Ontorio, Canada). Forehead and chin rests were used to minimize head movements. The spatial resolution of the eye tracker was 0.05°, and the sampling rate was 1,000 Hz. Participants used a Microsoft gaming button box to control the experiment and to answer comprehension questions.

2.3.Design

The experiment comprised three phases: pre-exposure, learning, and post-exposure. The core design was a 2 (test phase: pre-exposure vs. post-exposure) x 2 (spacing: distributed vs. massed) x 2 (contextual diversity: diverse vs. repeated) x 2 (retrieval: retrieval vs. no retrieval). Test phase, spacing and diversity were within-participant whereas retrieval practice was manipulated between-participants. Eye movements were monitored throughout the experiment, allowing us to compare reading behaviour pre-exposure vs. post-exposure, and during the learning phase itself. The dependent variables were first fixation duration, gaze duration, go past time and total time.

2.4.Materials

Our aim was to choose to-be-learned target words which would be unfamiliar to our participants at the start of the experiment. Forty-two rare English words (13 nouns, 17 adjectives and 12 verbs) were selected from norms provided by Brysbaert, Mandera, McCormick & Keuleers (2018) as being low prevalence words, where prevalence is derived from the percentage of people that know the word (M = -0.02; Range: -0.34-0.45) (see Keuleers, Stevens, Mandera & Brysbaert, 2015). Negative prevalence values indicate that words are not known by the majority of people, making them ideal stimuli for word learning experiments. Furthermore, three undergraduate students, who did not participate in the eye tracking experiment, confirmed that they did not know the 42 selected words. The words were also low in frequency (Zipf: M= 1.5; Range= 1.3-2.5; this is a standardized logtransformed measure of word frequency from SUBTLEX-US corpus; Brysbaert & New, 2009).

For each word, six different sentences were constructed: two neutral and four meaningful (see Table 1 for an example and Supplementary Materials for full list). Target words were always embedded in the middle of the sentences and sentence length averaged 71.8 characters and 13 words (Range= 38-129 characters; 6-21 words). The neutral sentences were not informative as to the meaning of the target word. These sentences were used to compare reading behaviour on the target word at baseline, before the learning phase, and afterwards, once the words had been seen in more meaningful sentences. For each word, one of its neutral sentences was used during the pre-exposure phase and the other during post-exposure, counterbalanced across participants.

The meaningful sentences provided some clues to infer the meaning of the target word and were read during the learning phase of the experiment. For target words in the contextually diverse condition, all four different sentences were read during the learning phase. For words in the repeated condition, only one sentence was read but it was repeated four times. For each participant, half of the words were presented in the diverse context condition and the other half were presented in the repeated context. The list of words presented in each condition was counterbalanced across participants.

During the pre-exposure phase and the learning phase, comprehension questions appeared on 25% trials. The purpose of these yes/no questions was to encourage participants to read for meaning, but they did not probe knowledge of the target word itself. At the end of the post-exposure test phase, a two-alternative forced choice comprehension question assessed how well the meaning of each target word had been learned.

Table 1. Example of the stimuli used in this experiment (example word *probity*)

	Diverse Context	**Repeated Context**
Neutral: pre-exposure	His family's probity was something that I noticed.	His family's probity was something that I noticed.
Exposure 1	This journalist prides himself on probity, but I actually don't think anything he wrote is true.	This journalist prides himself on probity, but I actually don't think anything he wrote is true.
Exposure 2	The boy who cried wolf lacked probity, so people eventually stopped believing him.	This journalist prides himself on probity, but I actually don't think anything he wrote is true.
Exposure 3	I can't name a single politician full of probity, they're all liars.	This journalist prides himself on probity, but I actually don't think anything he wrote is true.
Exposure 4	John's probity terrified her: he always told her exactly what he thought.	This journalist prides himself on probity, but I actually don't think anything he wrote is true.
Comprehension question	Do I think that the articles written by the journalist were true? **No** or Yes	
Neutral: post-exposure	A person who possesses probity is more likely to be successful.	A person who possesses probity is more likely to be successful.
Test question	What does probity mean? Sarcasm or **Honesty**	

2.5. Procedure

The experiment started with a calibration and validation procedure using a horizontal 3-point calibration. Calibration accuracy was always less than 0.25°; otherwise, calibration and validation were repeated. Participants were asked to read each sentence silently, and for comprehension. They were told that would be reading sentences that contained unknown words. They were instructed to read the sentences naturally and that they would see each unknown word several times so they should avoid spending extra time trying to memorize any unfamiliar words. Participants pressed a button on a Microsoft game controller to indicate they finished reading the sentence. This triggered presentation of the next sentence, or a comprehension question. The experiment lasted approximately 60 minutes and across the entire experiment, participants read two practice sentences followed by 252 experimental sentences. Drift measurements were taken at the beginning of each trial and calibration repeated when necessary.

Figure 1. Overview of the procedure

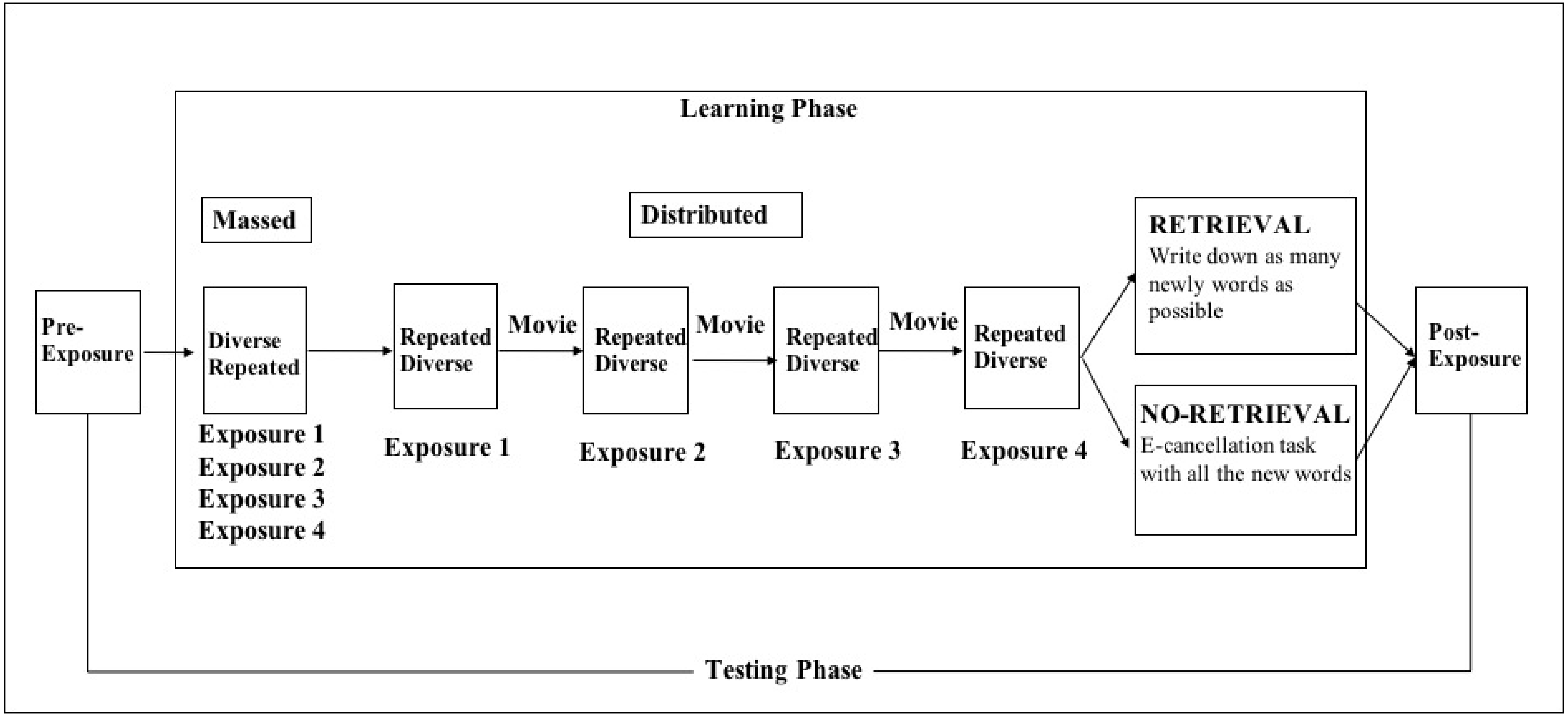


Figure 1 provides an overview of the procedure. The pre-exposure phase consisted of 42 neutral sentences, each containing one of the target words. Immediately following this was the learning phase, where the target words were experienced in meaningful sentences over five blocks of trials. To manipulate spacing, for each participant, half of the words were presented in the distributed condition (each occurring in one sentence in each four blocks) and half of the words in the massed condition (each occurring in four sentences, but all in one block only). In the distributed condition, the four blocks were separated by a break where participants watched a movie trailer lasting between 2 and 3 minutes. To avoid any confounding effects of order, the massed block appeared before the four distributed blocks for half of the participants, and after the distributed blocks for the other half.

To manipulate contextual diversity, half of the words were assigned to the diverse condition and the other half to the repeated condition, counterbalanced across participants and crossed with spacing. In the distributed condition, each word appeared in either the same meaningful sentence once in each of the four blocks (repeated context), or in a different meaningful sentence in each of the four blocks (diverse context). In the non-distributed condition, the meaningful sentences (either diverse or repeated) appeared in a random order within a single block, respecting the order of exposure. In both the pre-exposure and learning phases, comprehension questions occurred on 25% of trials. These did not tap knowledge of the target words but were intended to maintain attention and encourage reading for meaning.

The retrieval manipulation occurred immediately after the end of the learning phase. Participants in the retrieval group were asked to write down as many new words as they could remember. Those in the no-retrieval group experienced an additional exposure to each new word instead, via an e-cancellation task. They were presented with a piece of paper listing all of the target words, devoid of any sentence context, and asked to work through the list, striking all occurrences of the letter ‘e’.

Finally, the post-exposure phase consisted of 42 neutral sentences, each containing a target word. Note that for each word, a different neutral sentence was used to that seen during the pre-exposure phase. A two-alternative forced-choice comprehension question followed each sentence, directly probing how well the target word was understood.

2.6. Data Analyses

The "clean" function in DataViewer (SR Research) was used to trim the data. Fixations shorter than 80 ms, and which were located within one character space of the next or previous fixation, were merged into that nearby fixation; remaining fixations that were shorter than 80 ms or over 1200 ms were deleted.

The data were analysed by means of generalized linear mixed effects (glme) modelling using the glmer function from the lme4 package (Bates, Maechler & Bolker, 2012) within R (R Development Core Team, 2013)[1], using the gamma family and the identity link to reduce skew on row data (Lo & Andrews, 2015).

We considered four dependent variables: first fixation duration, gaze duration, go past time and total time. First fixation duration is the initial, first-pass fixation on a word, regardless of how many fixations it receives. Gaze duration is the sum of all consecutive first pass fixations on a word before leaving the word. Go past time is the time from the initial fixation on a word until the eyes move onward in the sentence; and total time is the sum of the duration of all the fixations on the target word, including regressions. Both first fixation duration and gaze duration are 'early' measures of processing time on a word, considered to reflect lexical processing, whereas go past time and total time are 'later' measures, typically

[1] Data and code for analyses are available in the Supplementary Materials.

though to reflect text integration processes (Rayner, 2009; see also Liversedge, Paterson & Pickering, 1998).

Specific details of the models used are provided in the Results section. Our general approach was to initially specify a full random structure for subjects and items, to avoid being anti-conservative (Barr, Levy, Scheepers & Tily, 2013). If a model failed to converge, we first increased the number of iterations using the optimizer "bobyqa" for glme models and then trimmed the random structure of the model until it converged (first by removing correlations between factors, and then interactions). All findings reported here are from models that converged successfully. Significance values and standard errors reflect both participant and item variability (Baayen, Davidson & Bates, 2008).

3. Results

Descriptive statistics for all phases of the experiment are summarised in Table 2. We first consider the pattern of effects found in the learning phase, as participants read meaningful sentences. We then go on to consider the reading of neutral sentences post-exposure, comparing against the baseline established in the pre-exposure phase.

3.1. Learning phase

A series of glme models were run to examine the effects of spacing (distributed vs. massed) and contextual diversity (diverse vs. repeated context) on each dependent variable during the learning phase (meaningful sentences), collapsing the data across the four exposures[2]. Note that during this phase of the experiment, the retrieval manipulation was yet to happen, so data from both retrieval groups were considered together. Spacing and

[2] As each word was experienced four times, we also examined reading behaviour as the learning phase unfolded across each exposure. However, models including exposure number (first, second, third and fourth) as a fixed factor failed to converge and thus aren't reported here.

contextual diversity were specified as categorical fixed factors. We used “contr.sdif” (MASS). Therefore, contrasts were specified as 0.5/-0.5 for the effects of spacing (distributed vs. massed) and contextual diversity (repeated vs. diverse), such that the intercept corresponds to the grand mean and the fixed effects to the main effect of the fixed factors. Both participants and items were specified as random factors. The results of the models are summarised in Table 3.

Table 2. Mean (SD) reading times as a function of group (retrieval vs. no-retrieval), spacing (distributed vs. massed) and contextual diversity (diverse vs. repeated context) in first fixation duration (FFD), gaze duration (GD), go past time and total time in the pre-exposure, learning and post-exposure phases of the experiment.

	Retrieval				**No-Retrieval**			
	Distributed		**Massed**		**Distributed**		**Massed**	
	Diverse	**Repeated**	**Diverse**	**Repeated**	**Diverse**	**Repeated**	**Diverse**	**Repeated**
FFD								
Pre-Exposure	259 (106)	262 (97)	255 (93)	253 (100)	249 (102)	240 (88)	239 (104)	243 (100)
Learning	239 (78)	230 (76)	237 (77)	232 (77)	232 (83)	224 (71)	231 (80)	218 (71)
Post-Exposure	249 (88)	243 (81)	233 (76)	244 (81)	223 (74)	229 (77)	229 (77)	227 (80)
GD								
Pre-Exposure	447 (279)	455 (298)	465 (272)	467 (302)	425 (268)	452 (313)	453 (338)	443 (284)
Learning	323 (166)	298 (154)	309 (148)	293 (145)	323 (191)	311 (208)	318 (195)	291 (184)
Post-Exposure	339 (173)	343 (191)	321 (149)	352 (213)	308 (151)	340 (198)	319 (204)	302 (164)
Go Past Time								
Pre-Exposure	611 (398)	586 (367)	606 (335)	604 (377)	533 (353)	563 (413)	609 (442)	563 (408)
Learning	392 (235)	346 (204)	372 (215)	340 (193)	388 (268)	354 (242)	376 (255)	339 (240)
Post-Exposure	395 (236)	412 (260)	390 (246)	442 (287)	353 (205)	391 (252)	365 (235)	371 (242)

Total Time								
Pre-Exposure	922 (511)	903 (542)	928 (525)	944 (526)	945 (773)	955 (734)	972 (777)	935 (722)
Learning	524 (358)	435 (311)	558 (393)	452 (333)	553 (441)	461 (388)	563 (484)	422 (330)
Post-Exposure	595 (376)	676 (417)	598 (355)	681 (437)	515 (378)	614 (467)	550 (432)	623 (433)

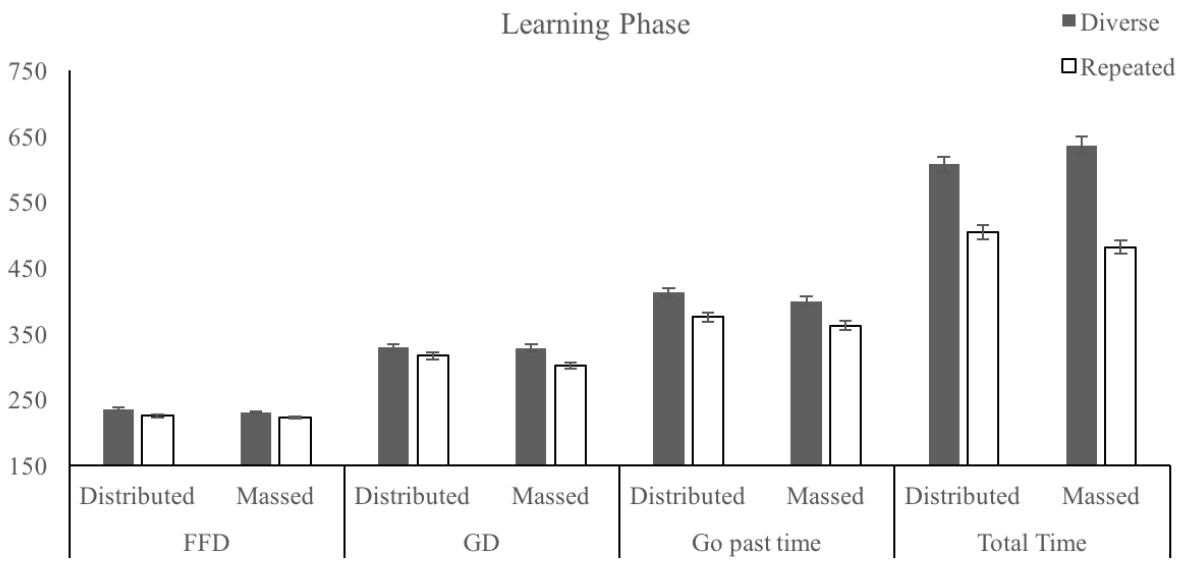
Table 3. Summary of the Glme model results from the learning phase, considering Spacing (Massed (Mass) vs. Distributed (Dist)) and Contextual Diversity (Repeated (Rept) vs. Diverse (Div)) as fixed factors for first fixation duration (FFD), gaze duration (GD), go past time and total time.

	FFD				GD				Go Past Time				Total Time			
	b	SE	*t*	p	*b*	SE	*t*	p	*b*	SE	*t*	p	*b*	SE	*t*	*p*
Intercept	233.12	3.16	73.83	0.00	314.03	2.84	110.64	0.00	370.43	3.26	113.46	0.00	501.06	2.83	177.13	0.00
Mass-Dist	-1.82	2.18	-0.83	0.40	-11.80	2.59	-4.55	0.00	-14.50	3.55	-4.08	0.00	1.41	2.45	0.58	0.56
Rept-Div	-9.15	2.51	-3.64	0.00	-21.35	2.73	-7.81	0.00	-37.84	2.96	-12.8	0.00	-110.7	3.63	-30.46	0.00
Mass-Dist: Rept-Div	0.27	3.02	0.09	0.93	-3.30	3.37	-0.98	0.33	5.24	3.32	1.58	0.11	-0.52	2.98	-0.18	0.86

Note: For FFD, GD and Go past time: glmer(depvar ~ Space*CD +(1+CD*Space|pp) + (1+CD*Space|stim), control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=10000)), data = Training, family= "Gamma"(link = "identity")); and for Total Time: glmer(depvar ~ Space*CD +(1+CD|pp) + (1+Space|stim), control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=10000)), data = Training, family= "Gamma"(link = "identity")).

There was an overall effect of contextual diversity in all dependent measures such that words seen in diverse contexts received longer fixations and longer reading times than words experienced in the same context (see Figure 2). Spacing influenced gaze duration and go past time, such that words experienced in the massed condition received shorter fixations and shorter reading times than words seen in the distributed condition. There were no interactions between spacing and contextual diversity. These findings suggest that during the learning phase, both contextual repetition and massed presentation allowed new words to be identified and integrated more easily, relative to when presentation was diverse and distributed.

Figure 2. Mean reading times and Standard Errors (error bars) as a function of contextual diversity (diverse vs repeated) and spacing (distributed vs. massed) during the learning phase.



3.2. Pre-exposure vs. post-exposure

All participants scored at least 75% on the comprehension questions asked throughout the pre-exposure and learning phases, indicating that they attended well and read for comprehension (M: 91%; SD: 4.8%). Performance levels were equivalently high in both retrieval groups (Retrieval: M: 90%, SD = 4.9%; No Retrieval: M = 92%, SD = 4.6%; p > 0.2). The mean score at post-test in response to comprehension questions tapping knowledge of the newly learned words was also high at 91% (SD: 6.3%). Once again, there was no difference between the two retrieval groups (Retrieval: M: 90%, SD = 7.0%; No Retrieval: M = 91%, SD = 5.5%; p > 0.3) or as a function of spacing and contextual diversity (distributed-diverse: M =90%, SD = 3.0%; distributed-repeated: M= 92%, SD = 3.0%; spaced-diverse: M= 90%, SD= 3.0%; spaced-repeated: M= 90%, SD= 3.0%).

We first ran a glme model to examine the effects of retrieval practice, spacing and contextual diversity on reading behaviour to the target words in neutral sentences, comparing pre vs. post-exposure performance. In this full model, variables were considered as categorical and specified as fixed factors, using “contr.sdif” (MASS). Therefore, contrasts were specified as 0.5/-0.5 and were used for the effects of group (retrieval vs. no retrieval), phase (pre vs. post exposure), spacing (distributed vs. massed) and contextual diversity (repeated vs. diverse), such that the intercept corresponds to the grand mean and the fixed effects to the main effect of the fixed factors. Both participants and items were specified as random factors. The results of the model are summarised in Table 4. We report results for first fixation duration, gaze duration and go past time; even a simplified model for total time failed to converge.

Table 4. Summary of the glme model results on neutral sentences, considering Phase (Pre-exposure (PreExp) vs. Post-exposure (PostExp)), Group (Retrieval (Ret) vs. No Retrieval (NoRet)), Spacing (Massed (Mass) vs. Distributed (Dist)) and Contextual Diversity (Repeated (Rept) vs. Diverse) as fixed factors, for first fixation duration (FFD), gaze duration (GD) and go past time.

	FFD				GD				Go Past Time			
	b	SE	*t*	p	*b*	SE	*t*	p	*b*	SE	*t*	p
Intercept (Global mean)	244.80	3.91	62.63	0.00	397.2	8.08	49.14	0.00	496.8	8.06	61.63	0.00
Ret-NoRet	14.20	5.07	2.80	0.00	15.81	6.51	2.43	0.01	26.13	10.29	2.54	0.01
PostExp-PreExp	-13.55	1.91	-7.08	0.00	-92.68	3.90	-23.76	0.00	-150.16	5.60	-26.81	0.00
Mass-Dist	-3.36	2.59	-1.29	0.20	1.63	7.28	0.22	0.82	14.10	6.44	2.19	0.03
Rept-Div	1.98	2.31	0.86	0.39	10.01	5.23	-1.91	0.05	10.18	4.90	2.08	0.04
Ret-NoRet:PostExp-PreExp	-0.58	3.44	-0.17	0.87	-6.11	5.59	-1.09	0.27	-38.24	6.42	-5.96	0.00
Ret-NoRet:Mass-Dist	-5.28	3.81	-1.38	0.17	11.98	8.89	1.35	0.18	-8.30	5.43	-1.53	0.13
PostExp-PreExp:Mass-Dist	3.73	3.42	1.09	0.27	-11.26	6.31	-1.79	0.07	-15.47	4.96	-3.12	0.00
Ret-NoRet: Rept-Div	-0.70	3.47	-0.20	0.84	-4.49	5.82	-0.77	0.44	-2.52	4.20	-0.60	0.55
PostExp-PreExp:Rept-Div	2.84	3.37	0.84	0.40	6.01	4.80	1.25	0.21	38.34	6.45	5.95	0.00
Mass-Dist:Rept-Div	2.92	3.31	0.88	0.38	-3.64	6.53	-0.56	0.58	4.76	5.05	0.94	0.35

Ret-NoRet:PostEx-PreExp:Mass-Dist	-5.77	4.85	-1.19	0.23	-15.99	6.77	-2.36	0.02	23.37	10.30	2.27	0.02
Ret-NoRet:PostExp-PreExp: Rept-Div	-1.69	4.61	-0.37	0.71	4.29	11.02	0.39	0.70	14.44	6.91	2.09	0.04
Ret-NoRet:Mass-Dist: Rept-Div	1.04	4.36	0.24	0.81	0.35	6.86	0.05	0.96	32.61	9.81	3.32	0.00
PostExp-PreExp:Mass-Dist: Rept-Div	2.11	4.36	0.48	0.63	12.31	6.12	2.01	0.04	21.41	5.46	3.92	0.00
PostExp-PreExp:Ret-NoRet: Mass-Dist: Rept-Div	37.90	6.53	5.80	0.00	45.88	7.34	6.25	0.00	-21.43	8.33	-2.57	0.01

Note: For FFD, GD and Go Past Time: glmer(depvar ~ Retrieval*phase*Space*CD + (1 +Space|pp) + (1+ CD|stim), control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=10000)), data = Testing, family= "Gamma"(link = "identity")).

There was an overall effect of learning such that participants made shorter fixation times in the post-exposure phase compared to the pre-exposure phase. Similarly, there was a main effect of retrieval group across all dependent measures, such that the participants who had the opportunity to retrieve the words made longer fixations and reading times than participants who had an extra exposure to the new words.

There was a four-way interaction between phase, retrieval, spacing and contextual diversity in first fixation, gaze duration and go past time. To simplify analyses, and given we did not expect any differences associated with yet-to-be manipulated factors in the pre-exposure phase, we ran a second set of glme models focusing on data from the post-exposure phase only. These assessed the effects of retrieval group, spacing and contextual diversity, with variables specified as fixed factors, using “contr.sdif” (MASS). Contrasts were specified as 0.5/-0.5 and were used for the effects of group (retrieval vs. no retrieval), spacing (distributed vs massed) and contextual diversity (repeated vs. diverse), such that the intercept corresponds to the grand mean and the fixed effects to the main effect of the fixed factors. Both participants and items were specified as random factors. The results of this model are summarised in Table 5.

Table 5. Summary of the glme model results on neutral sentences at post-exposure, considering Group (Retrieval (Ret) vs. No Retrieval (NoRet)), Spacing (Massed (Mass) vs. Distributed (Dist)) and Contextual Diversity (Repeated (Rept) vs. Diverse (Div)) as fixed factors, for first fixation duration (FFD), gaze duration (GD), go past time and total time.

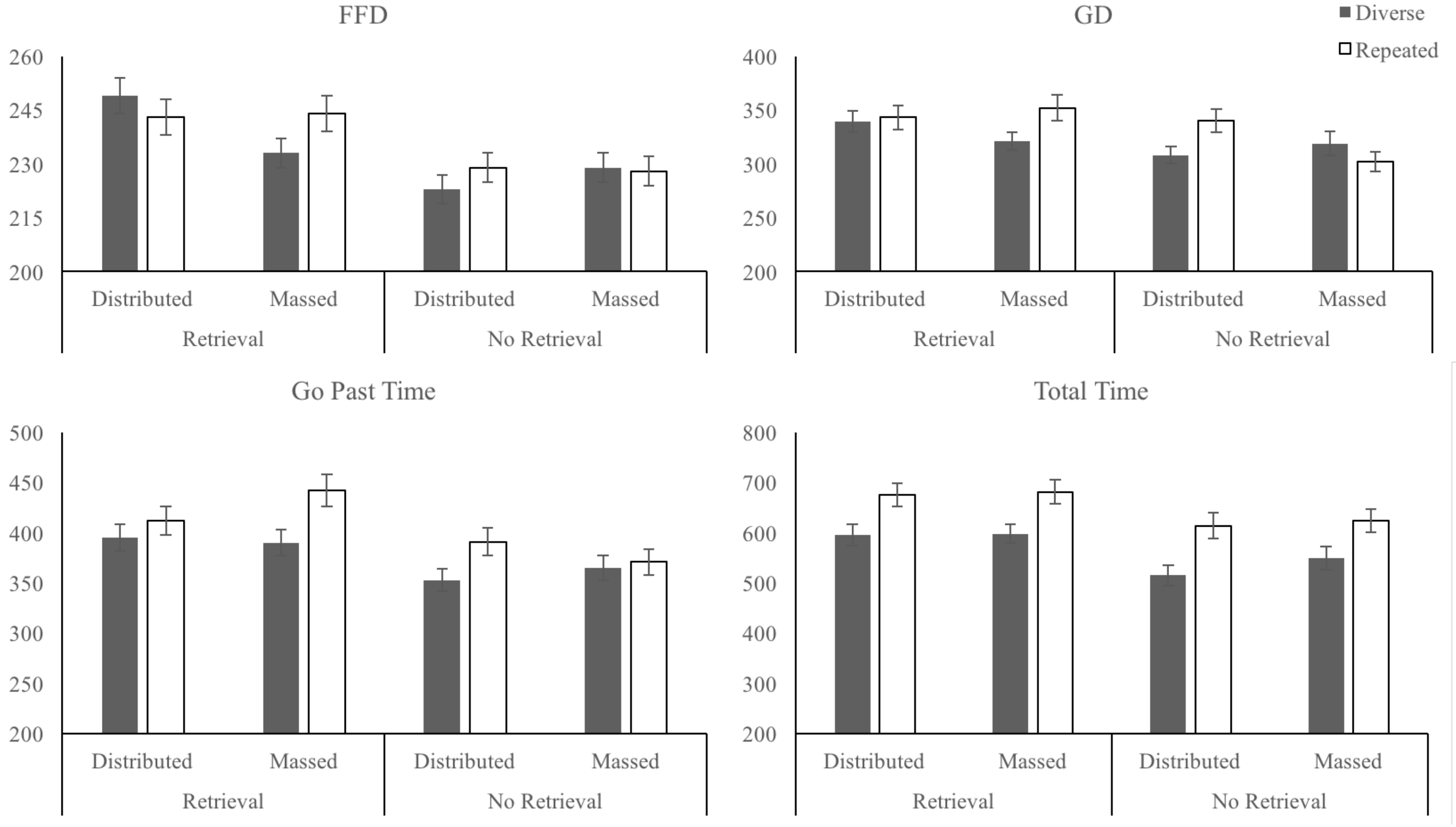
	FFD				GD				Go Past Time				Total Time		
	b	SE	*t*	p	*b*	SE	*t*	p	*b*	SE	*t*	p	*b*	SE	*t*
Intercept (Global mean)	237.24	5.09	46.59	0.00	340.00	10.74	31.66	0.00	402.91	8.67	46.49	0.00	625.78	7.17	87.28
Ret-NoRet	23.70	7.33	1.87	0.06	20.98	12.42	1.69	0.09	33.11	9.34	3.54	0.00	61.11	8.70	7.03
Mass-Dist	-2.39	3.16	-0.76	0.45	-7.42	5.44	-1.36	0.17	-1.03	4.76	-0.22	0.83	7.90	7.12	1.11
Rept-Div	2.14	3.19	0.67	0.50	12.15	4.88	2.49	0.01	27.62	6.46	4.27	0.00	89.36	6.88	12.99
Ret-NoRet: Mass-Dist	-7.37	4.63	-1.59	0.11	8.02	8.17	0.98	0.33	4.71	9.22	0.51	0.61	-30.98	7.10	-4.36
Ret-NoRet: Rept-Div	-2.02	5.08	-0.40	0.69	-1.21	7.70	-0.16	0.87	12.34	7.17	1.72	0.09	-19.88	7.78	-2.55
Mass-Dist: Rept-Div	3.79	4.51	0.84	0.40	-1.32	7.26	-0.18	0.86	5.13	8,20	0.62	0.53	-12.09	6.87	-1.76
Ret-NoRet:Mass-Dist: Rept-Div	18.82	6.49	2.90	0.00	24.66	8.38	2.94	0.00	11.72	9.23	1.27	0.20	-59.81	9.91	-6.04

Note: For FFD, GD and Total Time: glmer (depvar ~ Retrieval*Space*CD +(1+CD|pp) +(1+Space|stim), control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=10000)), data = PostTraining, family= "Gamma"(link = "identity")); for Go Past Time: glmer(depvar ~ Retrieval*Space*CD +(1+CD|pp) +(1|stim), control=glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=10000)), data = PostTraining, family= "Gamma"(link = "identity")).

There was an effect of contextual diversity in gaze duration, go past and total times indicating that words previously read in diverse contexts received shorter reading times than words that had been seen in repeating contexts. Note that this effect was in opposite direction to the one found in the learning phase. This suggests that words presented in repeated contexts were not learnt as well as words seen across in diverse contexts. Although contextual repetition provided an apparent advantage during the learning phase, our findings indicate that words experienced in this condition were not as well consolidated in memory, resulting in a processing cost when they appeared in a different (and neutral) context in the post-exposure phase. The main effect of spacing was not significant in any of the eye movement measures (ts < 1.36). There was however a main effect of retrieval in go past and total time, such that the participants who had the opportunity to retrieve the words made longer fixations and spent longer reading the target words than participants who had an extra exposure to them.

Complicating the interpretation of these main effects, there was a three-way interaction between group, spacing and contextual diversity for first fixation, gaze duration and total time (see Table 5 and Figure 3) and the nature of this interaction varied across the different dependent measures. We thus conducted tests of simple effects to investigate how spacing and contextual diversity influenced reading behaviour within each retrieval group of participants.

Figure 3. Means reading times and Standard errors (error bars) for for the three factors: group (retrieval vs. no-retrieval), contextual diversity (diverse vs repeated) and spacing (distributed vs. massed) in the post-training phase.



Focusing first on the group of participants in the retrieval condition, there was an effect of spacing on first fixation duration for words experienced in diverse contexts (*b* = 6.59, SE = 2.60, *t* = 2.53): diverse words in the massed condition had shorter fixation times than diverse words in the distributed condition. Later in processing, however, as tapped by total time, fixation times were shorter for diverse words in the distributed condition than those in the massed condition (*b* = -25.01, SE = 5.35, *t* = -5.66).

Turning to the participants who received an additional exposure rather than retrieval opportunity, there was an effect of spacing on total time spent reading diverse words (*b* = -30.32, SE = 5.35, *t* = -5.66), indicating that contextually diverse words that were also distributed over time received shorter fixation times than those in the massed condition. In addition, fixation times were shorter for non-diverse words experienced in the massed condition than the distributed condition, both for gaze duration (*b* = 8.79, SE = 4.13, *t* = 2.12) and total time (*b* = -15.84, SE = 5.77, *t* = -2.75).

Taken together, this pattern of effects indicates that having the opportunity to retrieve a newly experienced word impacts on its identification and integration into a new sentence context. For words appearing in different contexts and massed in time, an advantage of retrieval was observed early in processing (first fixation duration), associated with enhanced word identification. In contrast, when the diverse words were spaced across time, an effect of retrieval was observed only in the later stages of processing (total time), consistent with retrieval facilitating integration.

4. General Discussion

In this experiment, we investigated whether contextual diversity, temporal spacing and retrieval opportunity influenced how well adults learned and processed novel words encountered during silent reading. In the learning phase of the experiment, participants experienced each new word four times with both contextual diversity and temporal spacing being manipulated. We assessed learning from this exposure by comparing pre- vs. post-exposure phase data, as participants read the words embedded in neutral sentences. We also asked whether retrieval practice provided an additional boost to learning.

Overall, there was a clear effect of learning. People became familiar with the new words: they performed with high levels of accuracy on the post-test comprehension questions that probed knowledge of the new words. Similarly, they made shorter fixations and showed shorter viewing times on the new words in the post-training sentences compared to the pre-training sentences – an effect that was seen across all dependent measures. These findings are consistent with previous experiments that have examined word learning by measuring eye movements as people read (Bai *et al.*, 2013; Blythe *et al.*, 2012; Joseph *et al.*, 2014; Joseph & Nation, 2018; Liang *et al.*, 2015).

4.1.The influence of Contextual Diversity

Contextual diversity clearly influenced reading times during the learning phase. Words experienced in the same context had shorter viewing times than words experienced in diverse contexts. This effect appeared early in processing (first fixation and gaze durations) and was also visible in later measures of processing (go past and total time). In contrast to this processing advantage for words appearing in repeating contexts during learning, variability in contextual experience enhanced performance in the test phase. Words that had appeared in diverse contexts during learning had shorter gaze durations, go past and total times at test than words that had appeared in the same contexts. Thus, participants had less difficulty identifying and integrating the newly-learned diverse words in new contexts than words they had experienced equally often but in repeating contexts. It might be that contextual variation during learning allows the representations of words to become more context-independent and therefore more easily identified and retrieved and from this, more available to be integrated into new contexts (e.g., Bolger *et al.*, 2008). An alternative (or overlapping) possibility, derived from Perfetti's (2007) lexical quality hypothesis, is that contextual variability during learning enhances the quality of lexical-semantic representations and makes them easier to integrate into the new neutral sentence. Our data do not allow us to distinguish between these two possibilities. Further empirical work is needed, as well as more specification as to how features such as "high quality" and "context independence" should be defined, particularly in the context of reading words in meaningful text.

An important question to ask concerns the nature of contextual diversity. Following Bolger *et al.* (2008), we induced variation in contextual diversity by manipulating the number of unique contexts a word appeared in (one context repeated four times vs. four different contexts each seen once). As the number of exposures was matched across diverse and repeated conditions, our data show that this type of variability influences learning and processing beyond frequency. Importantly however, one limitation to our experiment concerns the use of verbatim repetition in the repeated condition. In natural language, contextual variability is not entirely repeated vs. entirely changing. Instead, there are gradations in contextual variability. Arguably, this is best tapped by metrics such as semantic diversity – a measure that captures the similarity in content of different contexts, not just the number of contexts. Semantic diversity is more closely associated with lexical processing than document count in tasks such as word naming and lexical decision (Hsiao & Nation, 2018; Jones, Johns & Recchia, 2012). Further work is needed to test whether semantic variability is at the root of the diversity effects seen in our experiment. It would be interesting to compare learning from episodes that vary in semantic content (see Joseph and Nation (2018) for initial findings from children). Furthermore, our experiment provides initial evidence to show that contextual variability influences word learning in a single session of silent reading; experiments are now needed that present contexts over time (and semantic variability in contexts over time) to capture in the laboratory the type of learning that happens via natural reading experience.

4.2.The influence of Temporal Spacing

In memory paradigms, there is good evidence that words presented under distributed practice are better remembered than those experienced in a massed condition (e.g., Bloom & Shuell, 1981; Cepeda *et al.*, 2006; Challis, 1993; Greene, 1989; Hintzman & Block, 1973; Verkoeijen *et al.*, 2004). Therefore, we predicted that words read in sentences in the distributed condition would receive shorter fixations than words experienced in sentences in the massed condition. In the learning phase, however, the opposite effect was seen: words experienced in the massed condition were identified more easily and had shorter re-reading times than words seen in the distributed condition.

It is important to note that the learning phase of our experiment provides a window on learning as it happens, as participants read new words and inferred their meaning from context. This is quite different to memory paradigms that tap the product of learning. It perhaps makes sense then that processing during the learning phase seems to be faster for words experienced in the massed condition. Faster processing during the learning phase might not be indicative of optimal learning however. To address this, we need to examine what happened in the post-exposure phase of the experiment. Here, while there was no advantage for spaced words, the apparent advantage for massed words seen in the learning phase was no longer evident on any eye movement measure. This pattern hints that a spacing advantage might emerge over time, consistent with the idea that massed presentation aids processing ‘in the moment’ but that ultimately, spacing leads to better learning and retention over time. Furthermore, our retention interval was short, with only a few minutes between the exposure and test phase; intervals of 24 hours or more are more typical in the spacing literature (e.g., Underwood & Ekstrand, 1967).

4.3.The influence of Retrieval

We predicted a positive effect of retrieval relative to an additional exposure. Overall however, we observed the opposite pattern of effects: participants who experienced an extra exposure via the e-cancellation task showed shorter reading times (in go past and total time) post-exposure than those participants who had a retrieval opportunity. This is inconsistent with the effects of testing and generation on memory (e.g., Barcroft, 2007; Roediger & Karpicke, 2006). There are three important caveats, however. First, retrieval practice occurred immediately after the study phase, arguably restricting the potential retrieval benefit we could observe in our experiment. Second, retrieval group interacted in complex ways for different eye movement measures and third, retrieval was manipulated between participants. Future research should examine and replicate these effects in a within-subject design, and with a longer interval between retrieval practice and test.

Focusing on the retrieval group only, our data showed differences in the early stages of processing words that had been experienced in diverse contexts and under massed practice, as tapped by first pass reading. In addition for these participants, words that were learnt in diverse contexts under distributed practice were better integrated into new contexts, relative to the same condition in the non-retrieval group. Although these effects were short-lived, it suggests that retrieval opportunity might bring qualitative differences to language processing, consistent with the idea that when people actively elaborate the information that they need to learn, they remember it more (e.g. Pyc & Rawson, 2009). As noted above, however, these findings must be interpreted cautiously given the complex pattern of interactions shown across the different eye movement measures, and the fact that different people participated in the retrieval vs. non-retrieval groups.

4.4. Implications for word learning via reading

Our experiment set out to explore how different types of variability influenced how well people processed and learned new words encountered during reading experience. We found that as people read new words in sentences, contextual diversity, spacing and retrieval opportunity all influenced online reading, both during the learning phase itself and at post-test. Clearly, to fully understand what is driving these effects, further research is needed to replicate and extend our findings. Nevertheless, we end by drawing two general conclusions, one methodological and one more theoretical.

In terms of methodology, our findings add to the evidence base showing the utility of eye movement paradigms to capture word learning. As noted earlier, text is complex and reading experience is vast. Investigating word learning via reading allows us to move closer towards understanding how new words are acquired in more naturalistic circumstances, from text (rather than word lists for example). The eye movement record itself provides a direct window on moment-by-moment processing, as it happens, rather than via a secondary task that has its own metacognitive load. This type of paradigm shows how processing encounters with new words impact on learning outcomes, and how this is influenced by features of reading experience. Future work should extend the approach further to look more complex texts and multi-session exposure as well as longer term learning gains.

In terms of theory, our findings demonstrate that the environment within which words are experienced matters for how they are processed, and how they are learned. As frequency was matched across conditions in the exposure phase, any differences in eye movement behavior cannot be a consequence of differences in frequency in any straightforward way.

Instead, they suggest that accounts of lexical processing during reading need to consider distributional information in the learning environment more broadly (Adelman, *et al.*, 2006; Jones *et al.*, 2018; Hsiao & Nation, 2018; Nation, 2017). More generally, this highlights the need to consider links between processing time (as a word is encountered) and developmental time (as words are accrued and dynamically updated with successive encounters over time): how a word is processed is a product of previous encounters with it, and each encounter contributes to long-term knowledge about that word, which then influences processing on future encounters. Our findings also fit with the view that models of reading and eye movement control during reading should implement the encoding and retrieval of memory traces as learning procedures, informed by theoretical accounts of learning and memory (e.g., Reichle & Perfetti, 2003). Finally, our findings also highlight the dynamic and incremental nature of word learning via reading experience and call for theoretical accounts of reading that take learning, context and memory into account.

5. Acknowledgements

We thank Helen Burridge, Antoni Czerwinski and Susanna Diver for research assistance and their help with creating the stimuli and with data collection. This work was supported by a grant from The Economic and Social Research Council (ES/M009998/1).

6. References

Adelman, Brown & Quesada (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17 (9), 814-823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4), 703.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, 2(6), 396-408.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412.

Bai, X., Liang, F., Blythe, H. I., Zang, C., Yan, G., & Liversedge, S. P. (2013). Interword spacing effects on the acquisition of new vocabulary for readers of Chinese as a second language. *Journal of Research in Reading*, 36 (S1).

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57(1), 35-56.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011.

Betts, H. N., Gilbert, R. A., Cai, Z., Okedara, Z. B., & Rodd, J. M. (2018). Retuning of lexical-semantic representations: Repetition and spacing effects in word-meaning priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44(7)*, 1130-1150. http://dx.doi.org/10.1037/xlm0000507.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45(2), 122-159.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, 2, 35-67.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, 74(4), 245-248.

Blythe, H. I., Liang, F., Zang, C., Wang, J., Yan, G., Bai, X., & Liversedge, S. P. (2012). Inserting spaces into Chinese text helps readers to learn new words: An eye movement study. *Journal of Memory and Language*, 67(2), 241-254.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4), 977-990.

Brysbaert, M., Mandera, P., McCormick, S.F. & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. Behaviour Research Methods,1-13. https://doi.org/10.3758/s13428-018-1077-9

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6), 1563.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354.

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(1), 225.

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2), 389.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology,*84(5), 795-805.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 1-26.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*, 66(3), 325-331.

Goldenberg, E. R., & Sandhofer, C. M. (2013). Same, varied, or both? Contextual support aids young children in generalizing category labels. *Journal of Experimental Child Psychology*, 115(1), 150-162.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3), 371.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, 16(2), 96-101.

Hintzman, D. L., & Block, R. A. (1973). Memory for the spacing of repetitions. *Journal of Experimental Psychology*, 99(1), 70.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, 10(5), 562-567.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, 30(1), 138-149.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review*, 23(4), 1214-1220.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. *Progress in Brain Research*, 232, 239–283. https://doi.org/10.1016/bs.plm.2017.03.008.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 66(2), 115. https://doi.org/10.1037/a0026727

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, 166, 190-211. https://doi.org/https://doi.org/10.1016/j.jecp.2017.08.010.

Joseph, H. S., 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(1), 238-248.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. The Quarterly Journal of Experimental Psychology, 68(8), 1665-1692. https://doi.org/10.1080/17470218.2015.1022560.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review*, 76(1), 82.

Liang, F., Blythe, H. I., Zang, C., Bai, X., Yan, G., & Liversedge, S. P. (2015). Positional character frequency and word spacing facilitate the acquisition of novel words during Chinese children's reading. *Journal of Cognitive Psychology*, 27(5), 594-608.

Liversedge, S.P., Paterson, K.B. & Pickering, M.J. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), Eye guidance in reading and scene perception (pp. 55-75). Oxford: Elsevier Science Ltd.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Frontiers in Psychology, 6, 1171.

Lowell, R., & Morris, R. K. (2014). Word length effects on novel words: Evidence from eye movements. *Attention, Perception, & Psychophysics*, 76(1), 179-189.

McDaniel, M. A., & Mason, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 371-385.

Nagy, W.E., Herman, P.A. & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233-253

Nation, K. (2017). Nurturing a lexical legacy: reading experience is critical for the development of word reading skill. *npj Science of Learning*, 2(1), 3. https://doi.org/10.1038/s41539-017-

0004-7.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29(4), 559-586.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2), 101.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357-383.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1), 275.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437-447.

Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, 27(3), 431-452.

Rashotte, C. A., Torgesen, J. K. & Wagner, R. K. (1999). TOWRE: Test of word reading efficiency. Austin, TX: Pro-ed.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457-1506.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, 22(4), 425-469.

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.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20-27.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249-255.

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(6), 1582.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558-1568.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 6(2), 175-184.

Twomey, K. E., & Westermann, G. (2018). Learned Labels Shape Pre-speech Infants’ Object Representations. *Infancy*, 23(1), 61-73.

Underwood, B. J., & Ekstrand, B. R. (1967). Effect of distributed practice on paired-associate learning. *Journal of Experimental Psychology*, 73(4p2), 1.

Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4), 796.