Anchoring and contextual variation in the early stages of incidental word learning during reading

Matthew HC Mak, Yaling Hsiao, Kate Nation

Department of Experimental Psychology, University of Oxford

Author Note

We have no conflict of interest to disclose. Correspondence concerning this article should be addressed to Matthew Mak, Department of Experimental Psychology, University of Oxford, United Kingdom; +44 1865281668; matthew.mak@psy.ox.ac.uk

Acknowledgements

Data in this article were presented at the 2020 EPS January meeting in London, UK, and at the 2019 LKALE SIG conference in Oxford, UK. This research was supported by the R C Lee Centenary Scholarship to Matthew Mak and an Economic and Social Research Council grant (ES/M009998/1) to Kate Nation. Yaling Hsiao is supported by a British Academy Post-Doctoral Fellowship (PF2/180013). We thank the three reviewers for their comments on an earlier draft of this paper.

Open Science Statement

All data, scripts, materials, and simulation procedure are publicly available on Open Science Framework (https://osf.io/sjwgr/).

Abstract

Lexical processing is influenced by a word's semantic diversity, as estimated by corpus-derived metrics. Although this suggests that contextual variation shapes verbal learning and memory, it is not clear what semantic diversity represents and why this influences lexical processing. Word learning experiments and simulations offer an opportunity to manipulate contextual variation directly and measure the effects on processing. In Experiment 1, adults read novel words in six naturalistic passages spanning one familiar topic (low semantic diversity) or six familiar topics (high semantic diversity). Words experienced in the low-diversity condition showed better learning, an effect replicated by simulating spreading activation in lexical networks differing in semantic diversity. We attributed these findings to "anchoring", a process of stabilizing novel word representations by securing them onto a familiar topic in long-term memory. Simulation 2 and Experiment 2 tested whether word learning might be better placed to take advantage of diversity if novel words were given an anchoring opportunity. Simulations and behavioural data both showed that after an anchoring opportunity, novel words forms were better learned in the high-diversity condition, contrasting with Simulation/Experiment 1. Taken together, these findings show that anchoring and contextual variation both influence the early stages of word learning.

Keywords: word learning, contextual variation, semantic diversity, anchoring, network simulation

Introduction

Some psycholinguistic theories assert that the mental representations of words comprise not only the words themselves, but also the linguistic contexts in which they have been experienced (Bolger, Balass, Landen, & Perfetti, 2008; Perfetti, 2007; see also Nelson & Shiffrin, 2013). This entails that each word has its unique contextual history, shaped by a person's language and reading experience (Nation, 2017). Converging evidence from behavioural and computational studies shows that the contextual history of a word influences the efficiency with which it is learned, remembered and processed in lexical tasks (for review, see Jones, Dye, & Johns, 2017). In this paper, we investigated how the incidental learning of novel words is influenced by their respective contextual history.

How best to capture contextual history is an open question. One approach is to consider the number of unique documents a word appears in across a corpus, sometimes termed contextual diversity (e.g., Adelman, Brown, & Quesada, 2006). While associated with lexical processing, contextual diversity is tightly correlated with frequency leading to the view that it might be a more reliable and valid index of frequency, rather than a reflection of the content of word's contextual history (e.g., Brysbaert & New, 2009; see also Hollis, 2020). Another approach is to use latent semantic analysis to derive a metric known as semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013). A word's semantic diversity value corresponds to the degree of difference in semantic content between all the contexts in which the word appears across a corpus. For example, a word like *spinach* is low in semantic diversity: It tends to occur in a limited range of contexts, most of which are food-related (Hoffman & Woollams, 2015). In contrast, a word like *chance* appears in all sorts of contexts, making it high in semantic diversity. Put simply, *chance* provides little to no clues as to the content of a context whereas *spinach* provides a reasonable clue. High-diversity words are easier to identify in visual word recognition tasks such as lexical decision (e.g., Hoffman et al., 2013; Hoffman & Woollams, 2015; Hsiao, Bird, Norris, Pagán, & Nation, 2019; Hsiao & Nation, 2018; for a similar approach see Jones, Johns, & Recchia, 2012; Johns, Sheppard, Jones, & Taler, 2016). This facilitative effect is not accounted for by other variables such as frequency, contextual diversity, or age-of-acquisition.

Why are words that have been experienced in more diverse contexts easier to identify in lexical decision? To address this question, Jones et al. (2012) developed the Semantic Distinctiveness Model. In this distributed model, when a word is experienced in text, this encounter will be compared to the word's contextual history, stored in its memory vector. If the new context is distinct from the word's stored contextual history, this will exert pressure on the model to update the word's representation. In contrast, if the context is similar to what is stored in the memory

vector, less updating occurs. The extent to which a word is updated on each encounter depends therefore on contextual change. Over time, those words experienced in more diverse contexts may become more context-independent (i.e., less associated with a specific context), and as a by-product, they should become easier to identify. At the same time, the Semantic Distinctiveness Model predicts that variation in contextual experience should make high-diversity words harder to process in terms of meaning. In line with this prediction, the same words that show a processing advantage in visual lexical decision are disadvantaged in tasks that require semantic judgements to be made (Hoffman & Woollams, 2015; Hsiao et al., 2019).

Another way to account for the effect of semantic diversity on word recognition and semantic judgement is to draw a parallel with the effect of polysemy (i.e., words with multiple distinct but related meanings). Polysemous (vs. non-polysemous) words are recognized faster in lexical decision but are judged less efficiently in tasks requiring semantic judgement (e.g., Hino, Pexman, Lupkar, 2006; Rodd, Gaskell, & Marslen-Wilson, 2002; 2004; see Rodd, 2020 for review), mirroring the effects of semantic diversity. Polysemy tends to be classified categorically, based on the number of discrete definitions it has (e.g., Jastrzembski, 1981). Semantic diversity, in contrast, defines polysemy based on the notion that polysemous variation in word meaning is a continuous, graded property, which is a product of variation in the context in which the word is used (Hoffman & Woollams, 2015; Hoffman et al., 2013; but see Cevoli, Watkins, & Rastle, 2020).

Although the findings reviewed thus far are consistent in pattern, there are three important limitations and gaps in knowledge. First, semantic diversity correlates moderately with a range of other lexical variables, including frequency (r = .49, Hoffman et al., 2013), age-of-acquisition (r = -.40, Hsiao & Nation, 2018), concreteness (r = -.51, Hoffman et al., 2013), and number of senses (r = .36, Hoffman et al., 2013). While it is possible to regress these factors out in statistical analyses, they might well be driving the semantic diversity effect. Second and somewhat relatedly, while something about item-level contextual history appears likely to be involved in shaping a word's semantic diversity value, it is not clear exactly what it is tapping. For example, Cevoli et al. (2020) suggested that the term "textual diversity" may be a more appropriate description of the underlying construct. Finally, a person's contextual experience with a word is unique, depending on factors such as education level; however, a word's semantic diversity value is derived from a language corpus containing many thousands of texts (e.g., British National Corpus), which does not necessarily reflect the person's language experience. Relatedly, a word's diversity value varies significantly depending on what type of corpus is used (e.g., spoken vs. written corpus, children vs. adult literature; Hsiao & Nation, 2018; Johns, Dye, & Jones, 2020). Taken together, it is clear that while semantic diversity captures variance in lexical processing, many questions remain about its nature. One way to address

all these questions is via word-learning experiments where new words are encountered in different contexts and the effects of this on future processing is assessed. This methodology allows precise control of lexical statistics (e.g., frequency) and the nature of the contextual encounters. It also offers the potential to track learning over time, as knowledge grows with successive encounters.

A small number of studies have taken this general approach. Unfortunately however, and in contrast to the consistent evidence base emerging from studies using corpus-derived semantic diversity metrics, the pattern of results is not clear or consistent across studies. This is perhaps not surprising as each induced diversity in different ways, indexed learning using different tasks, and used different participant groups (undergraduate students vs. school children). Nevertheless, it is useful to review them briefly.

Johns, Dye, and Jones (2016) had undergraduates read pseudowords (e.g., *covella*) embedded in same-themed (low semantic diversity) or different-themed (high semantic diversity) passages. Post-exposure, participants completed a speeded recognition task (akin to lexical decision) where they judged whether each item had been seen in any of the passages. They found that while newly-learned words experienced in the different-themed passages were identified more efficiently than those experienced in the same-themed passages, they were processed less well in a semantic similarity judgement task. These behavioural observations fitted with simulations from the Semantic Distinctiveness Model reported by Johns et al. (2016). Rosa, Tapia, and Perea (2017) also found a diversity advantage in learning. They asked 8-9 year-olds to read unfamiliar words embedded in either same-themed or different-themed texts. Afterwards, words experienced in different-themed texts were processed more effectively in a range of tasks. Like Johns et al. (2016) there was a diversity advantage on tasks tapping word recognition; unlike Johns et al., however, there was also a diversity advantage on tasks tapping knowledge of word meaning. In contrast to these findings, Joseph and Nation (2018) reported no difference in learning outcomes for words encountered in high vs. low diversity contexts; nor were there any differences in online processing, as indexed by the children's eye movements.

In summary, item-level contextual history as captured by a corpus-derived semantic diversity metric affects lexical processing (e.g., Hoffman & Woollams, 2015), but there is a clear need for data from learning experiments to understand its nature. Yet, the existing evidence base is small, methodologically varied and contradictory in nature. With this as a backdrop, we conducted two word learning experiments with young adults, both similar in method to Johns et al. (2016) but considering in more detail topic familiarity, as discussed below. To complement these behavioural experiments we ran two computational investigations to examine the effect of semantic diversity by simulating spreading activation in lexical networks.

Experiment 1

Following Johns et al. (2016), young adults read passages containing novel words; there was no instruction to learn the words as we aimed to capture incidental learning. In one condition, passages were set around the same theme (low semantic diversity). Word learning in this condition was compared with learning following exposure to words in different-themed passages (high semantic diversity).

Our experiment differed from Johns et al. in a number of ways. We increased the familiarity of the topics included in the exposure phase. Johns et al. used relatively obscure topics including passages about Strauss-Kahn (a French politician), Aundhati Roy (an Indian writer), Braybrooke (a theologian), Moore's Law, and Sunni Islam, whereas ours contained more familiar topics such as Brexit, David Bowie, and Donald Trump. We did this as pilot work indicated that topic familiarity influenced word learning (Mak & Nation, 2019; see also Havas et al., 2018); we were also keen to motivate participants to read for interest and capture word learning via regular reading experience, rather than active study. Johns et al. measured learning using speeded recognition and semantic relatedness judgement. We added a third measure, namely a pseudoword inferiority task.

As this task has not been used to index word learning before, it is useful to set out our rationale here. The pseudoword inferiority effect, built upon the traditional word superiority effect (Reicher-Wheeler task; Reicher, 1969; Wheeler, 1970), refers to the phenomenon that certain letters are recognised more poorly when they are seen in a pronounceable pseudoword (e.g., B**E**NANA) relative to a string of Xs (e.g., X**E**XXXX). To illustrate, consider the pseudoword BENANA, which differs from the familiar word BANANA in terms of the second letter. BENANA is shown on the computer screen for a short time (e.g., 50 ms) and participants are then asked to indicate whether it was an E or an A that they saw in the second letter position. Grainger and Jacobs (2005) reported that letter identification is poor in pseudowords such as BENANA: participants tended to choose A more than E, relative to the control condition when the target letter appeared in the same letter position but in a string of Xs (e.g., XEXXXX). This phenomenon, known as the pseudoword inferiority effect, is presumably due to top-down interference from familiar words (in this case, BANANA).

Extending this rationale to capture word learning, if the form of a novel word has been well-established, it should induce a pseudoword inferiority effect. For example, if a participant experienced a novel word such as OSSANIC while reading, and it has gained word-like status (e.g., being well-specified), it should then bias letter recognition for a related pseudoword such as OSSA**M**IC in this letter identification task; that is, if OSSA**M**IC is shown briefly followed by a forced-choice decision of N or M at the fifth position, having learned OSSANIC should bias false recognition

of N, relative to a baseline condition of XXXXMXX. If, however, OSSANIC had not been learned well during incidental reading, less bias should be seen. In Experiment 1, we asked whether the pseudoword inferiority effect is sensitive to word learning as induced by variations in contextual history.

Method

Participants

Forty-five adults from the Oxford University participant pool (13 males; M_{age} = 22.53; SD_{age} = 4.68) participated in this experiment for either course credits or payment (£10/hour). All were native English speakers with normal or corrected-to-normal eyesight and none reported history of any developmental disorders. Five participants were excluded: two because their accuracy rates in the comprehension question were 2 SDs below the sample mean, and the other three because their accuracy rates in speeded recognition were 2 SDs below the sample mean. All analyses reported here, unless otherwise specified, are based on the remaining 40 participants.

Materials and Procedure

The experiment comprised two phases, (i) an exposure phase where variation in semantic diversity was induced and (ii) a surprise test phase that measured knowledge and processing of the newly learned words. Both phases were programmed and run using Gorilla (www.gorilla.sc), an internet-based platform. Participants completed both phases in a single 45-minute session held in a computer laboratory; up to five participants were tested at a time, well-spaced in the room.

(i) Exposure phase

Ten relatively rare English words were chosen as base words (e.g., *mendacious*, *predilection*; mean SUBTLEX frequency: 0.454/million; full list in Appendix A). Twelve passages were constructed for each base word: six sampled the same discourse topic and formed the low semantic diversity condition; the other six spanned distinct topics, forming the high semantic diversity condition. The topics were chosen to be familiar to an average undergraduate in the UK; examples included Brexit, David Bowie, and Donald Trump. The passages were excerpted and edited from respectable non-fiction sources (e.g., The New York Times, The Guardian). Table 1 shows a sample passage.

Latent Semantic Analysis confirmed that the degree of semantic overlap between the passages was significantly higher in the low-diversity condition (M = 0.49) than in the high-diversity condition (M = 0.19), $t(10) = 3.39$, $p = .006$. Passages in the two conditions were matched for the

number of words (M_{Low}= 125.6; M_{High} = 128.5), $t(71) = -0.85$, $p = .394$, and readability, as indexed by Flesch Reading Ease test, $t(71) = 0.40$, $p = .691$.

Our next step was to replace the 10 base words with pronounceable pseudowords (e.g., OSSANIC, LONBUST, APPURATE). Two sets of ten pseudowords were chosen from the English Lexicon Project (Balota et al., 2007) and matched on a range of lexical properties, including word length, number of orthographic neighbours, and bigram frequency (Appendix B). Each participant encountered pseudowords from one of the two sets, and assignment to pseudoword set was randomized.

Each participant encountered ten pseudowords, five in the high diversity condition and five in the low diversity condition. In other words, semantic diversity was manipulated within-subject. Assignment to the semantic diversity condition was fully counterbalanced on the item level. Participants read each item six times with frequency matched exactly across diversity conditions. They were naïve to the purposes of the experiment and were told that they would be reading English passages from a standardized test designed for non-native speakers of English. Note that the register of the passages was relatively formal to align with the fact that in real life, unknown or rare words are typically encountered in more formal or academic texts. Our aim was to reduce the likelihood of participants deducing that the target words were fake replacements for existing words.

A reading trial began with a fixation cross of 600 ms followed by a passage. A Likert scale appeared below the passage after 15 seconds. The participants rated how comprehensible the passage was on a continuous scale of 1 (No understanding) to 7 (Perfect understanding). They had a maximum of 60 seconds to read and rate each passage. Reading time was recorded. After the rating was submitted (or after 60 seconds had elapsed), the passage and scale were replaced by two comprehension questions; these served to encourage reading for meaning. The question format was a mixture of multiple-choice, fill in the blank, and true-false. They were straightforward and did not relate to the target words. Pilot testing confirmed that question difficulty was equivalent across diversity conditions.

Each participant read a total of 36 unique passages, divided equally into six blocks. Block order was randomized, as was the trial order within each block. In summary, each participant experienced five pseudowords in the low-diversity condition and five in the high-diversity condition. Regardless of the assigned condition, each pseudoword was seen a total of six times in the exposure phase.

Table 1
A sample passage in Experiment 1.

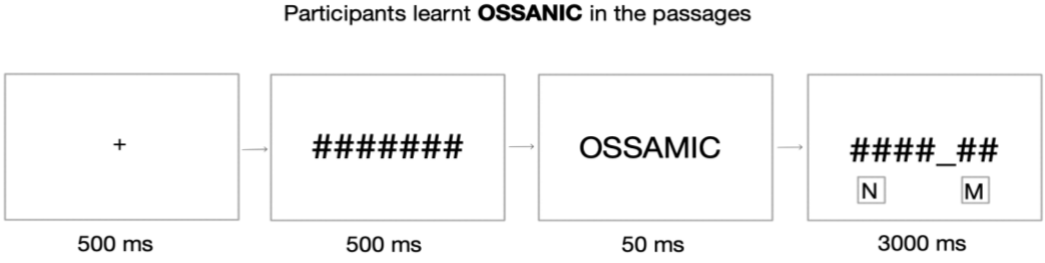
Base word	mendacious
Replaced by	ossanic [50% participants] *OR* chrabble [50% participants]
Passage	Though Trump is quick to deride unfavourable news stories as 'fake news', the administration owes its existence to the way mainstream media platforms can accommodate an avalanche of stories. Whether negative or fawning, any media attention popularize Trump's fact-free announcements and racist tweets. Through the sheer audacity of his ossanic personality, amplified a million-fold by his occupancy of the most powerful office on the globe, Trump is a one-man train-wreck. And, with each jarring twist and turn, the nation's media dutifully conveys his evermore crazy mutterings. While those in the mainstream media like to portray themselves as impartial observers, which at times forces them to take an antagonistic position towards the White House's shenanigans, more often than not reporters and editors are more like trumpets being played by a blow-hard president.
Note	Target words were **not** underlined in any of the passages that participants read.

(ii) Test phase

Immediately following the exposure phase, participants completed three post-tests in a fixed order, namely pseudoword inferiority, speeded recognition, and semantic relatedness judgement.

Pseudoword inferiority. This was modelled on the Reicher-Wheeler task (see Figure 1). A trial started with a fixation cross for 500 ms, followed by a 500 ms mask. Next, a string of letters appeared for 50 ms. This was then replaced by a string of hashtags surrounding an underscored blank space. Participants were asked to select (using the mouse) which letter had appeared in the blank space, choosing from the two alternatives presented underneath the stimulus. The next trial started automatically once a response was recorded or after 3000 ms had elapsed.

Figure 1
A sample trial in the pseudoword inferiority task, Experiment 1.



There were three conditions in this task: novel word, X, and real word. For the novel word condition, stimuli were derived from the 10 newly learned pseudowords but deviating by one letter (e.g., OSSA**N**IC became OSSA**M**IC). The deviating letter was always the letter replaced by an underscore. In this example, the two alternative choices were M (the correct answer) and an N (incorrect answer, but aligning with the recently learned item). To increase power, there were two trials for each of the newly learned words, each with a different deviating letter (e.g., for OSSANIC, test stimuli were OSSA**M**IC and OSS**E**NIC). This increased the number of observations per diversity condition from five to 10 for each participant.

Participants also saw each deviating pseudoword in the X condition. For example, if OSSANIC was encountered during the exposure phase, they would see OSSAMIC and XXXMXX in this task, with M as the correct answer and N as the incorrect answer.

The real word condition comprised 18 medium-frequency words (e.g., FIREWORK; mean SUBTLEX frequency: 10.5/million). One letter was changed to make it a pronounceable pseudoword (e.g., FIREWORK → FIREWARK). Participants chose between the deviating letter (e.g. A, the correct answer) and the letter that would produce the real word (e.g. O, the incorrect answer). Items in the condition served as the positive control.

There were a total of 56 experimental trials, whose order was randomised. The task started with five practice trials, where familiar words (e.g., DESKTOP, LAMP) were shown.

Speeded recognition. Modelled on a standard lexical decision paradigm, participants saw a pseudoword centred on the screen and were asked to determine whether they had encountered that in the reading passages, responding as quickly and as accurately as possible.

Following two practice trials, each trial began with a fixation cross displayed for 800 ms, followed by a stimulus for 1800 ms, during which the participants pressed buttons to make a judgement. For half of the participants, the A key was used for “Yes: I have encountered this word” and the L key for “No: I have not encountered this word”; the assignment of keys was reversed for the other participants. The next trial started as soon as a response was recorded, or if no response was recorded after 1800 ms. No feedback was given. Both accuracy and RTs were recorded.

Of the 20 pseudowords selected for the experiment (see exposure phase), each participant saw 10 of them in the reading passages; these were presented for speeded recognition, along with the other 10 unseen pseudowords, which served as foils here. The 20 items were presented in a random order in a single block. The block was repeated three times in total with a random order each time. Overall, each participant completed 60 trials: 30 ‘yes’ items (the 10 newly learned pseudowords x 3) and 30 ‘no’ items (the 10 unseen pseudowords x 3).

Semantic relatedness judgement. In each trial, participants saw one of the 10 newly learned pseudowords on the left side of the computer screen (e.g., OSSANIC) and a familiar English word (e.g., DISHONEST) on the right side. Their task was to judge, as quickly and as accurately as possible, whether the two words were related in meaning. Each pseudoword was seen on four occasions, twice paired with an associated word ('related' trials) and twice with an unrelated word ('unrelated' trials).

To generate the associated words, three native speakers of English who did not participate in the main study were asked to give three semantic associates for each of the 10 base words (e.g., *mendacious*). From these responses, the two most popular ones were selected, none of which appeared in the passages (e.g., for *mendacious*, the associates were *dishonest* and *fraudulent*; see Appendix C for complete list). To create the unrelated trials, each pseudoword was paired with two different words unrelated in meaning, both randomly taken from the list of associates.

Each trial started with a fixation cross for 500 ms, then two words appeared on the screen. Participants had 2500 ms to decide if they were related in meaning. Stimuli were presented in black, uppercase courier-new font on a white background. For half of the participants, the left arrow key was used for "They are related" and the right arrow key for "They are unrelated"; the assignment of keys was reversed for the other participants. After a response, or if no response was recorded after 2500 ms, the next trial started automatically. Both accuracy and RTs were recorded. Each participant completed 40 trials in total, divided into two blocks. Block order was randomised, as was the trial order within each block. This task started with a practice block containing 10 trials with familiar English words as stimuli (e.g., EYE——EAR).

Results

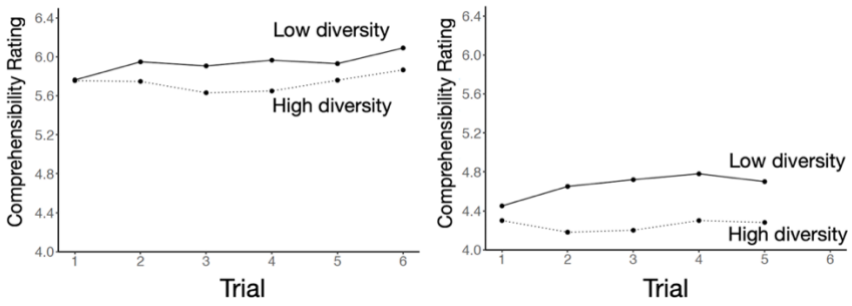
(i) Exposure phase

Performance on the comprehension questions was satisfactory (M = 84.5%, SD = 5.96%) indicating that the passages were read for meaning. Comprehension was equivalent across the semantic diversity conditions, $t(39) = 0.12$, $p = .76$. There was however a diversity effect on passage reading time, with passages in the high-diversity condition (M = 36.05 sec, SD = 10.3 sec) taking longer to read than those in the low-diversity condition (M = 34.79 sec, SD = 9.9 sec), $t(39) = 3.23$, $p = .002$. While all passages received a high comprehensibility rating, those in the low-diversity condition were rated as more comprehensible (M = 5.94, SD = 0.96) than those in the high-diversity condition (M = 5.74, SD = 0.99), $t(39) = 4.58$, $p = < .001$. This suggests that subjective comprehension of the low-diversity passages (all on the same topic) grew as participants become more familiar with that topic. Note Johns et al. (2016) reported an average ratings of about 4.5 (out of 7); as intended,

therefore, our passages were more comprehensible (see Figure 2 for comparison), probably due to increased topic familiarity.

Figure 2

Comprehensibility ratings for high and low diversity passages across trials in Experiment 1 (left) and Johns et al. (right). Higher ratings (max = 7) indicate that passages in Experiment 1 were easier to comprehend.



(ii) Test phase

Analytic approach

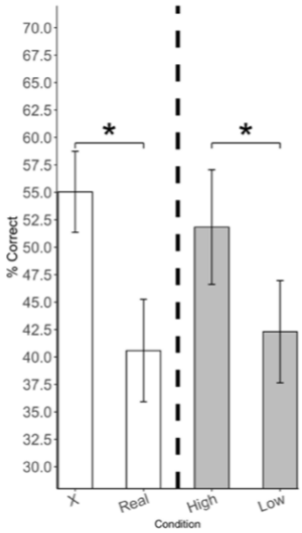
All data were analysed in a mixed-effects environment using the lme4 package (version 1.1.17; Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2019). Generalised linear mixed-effect models were used for accuracy data while RT data were analysed using linear mixed-effect models. RT data were transformed using either inverse or log transformation; the one that resulted in more normal residual distribution, as indexed by qqPlot (Millard, 2013), was adopted. All the models included random effects for participants and items, and a fixed effect for semantic diversity, which had two levels (low diversity vs. high diversity). This was coded using sum contrasts. Following Barr et al. (2013), models were first computed with a maximal random structure. If a model failed to converge or produced a singular fit, the random-effect structure was simplified (e.g., Barr et al., 2013; Bolker et al., 2009). The explanatory value of the converged model was compared with a random-intercept-only model, using anova model comparison. If the two models did not differ (as indexed by a threshold of chisq < 3; Jaeger, 2011; see also Gaskell, Cairney, & Rodd, 2019), the random-intercept-only model was favoured as this is considered more statistically powerful (Matuschek et al., 2017). However, if the two models differed (chisq > 3), the conservative model with the maximal (or the near-maximal) structure was favoured and reported (all data and scripts are available on OSF).

Pseudoword inferiority. Accuracy across the four conditions is plotted in Figure 3. The first point to highlight is that the basic pseudoword inferiority effect was present, replicating earlier findings (Grainger & Jacobs, 2005): Divergent letters embedded in real words (e.g., the A in

FIREWARK) were recognised more poorly (*M* = 40.5%) than letters embedded in a string of Xs (*M* = 55%), *z* = 2.43, *p* = .015.

Figure 3

Mean accuracy rates across the four conditions in the pseudoword inferiority task, Experiment 1.



Note. (i) Lower performance in the real word and low conditions is indicative of greater top-down interference. (ii) Error bars represent 95% within-subject confidence intervals.

Turning to the newly learned words experienced in the exposure phase (grey bars in Figure 3), letter identification was significantly poorer in the low-diversity condition (*M* = 42.8%) relative to the high-diversity condition (*M* = 51.8%; *z* = -2.20, *p* = .028, see Table 2 for model summary). This suggests that the low-diversity passages allowed the novel words to establish stronger and/or more well-specified lexical representations, leading to greater error in this letter identification task.

Table 2

*Model summary of the GLME models examining the effect of semantic diversity on **accuracy** in each of the post-tests, Experiment 1.*

	Estimate	Standard Error	z value	p value
(a) Pseudoword inferiority				
Intercept	-0.20	0.21	-0.98	.325
Sem. Diversity	0.18	0.08	2.20	.028*
(b) Speeded recognition				
Intercept	3.48	0.42	8.30	<.001*
Sem. Diversity	0.67	0.31	2.12	.034*
(c) Semantic relatedness				
Intercept	0.66	0.12	5.48	<.001*
Sem. Diversity	0.40	0.06	6.60	<.001*

Table 3

*Model summary of the LME models examining the effect of semantic diversity on **RTs** in each of the post-tests, Experiment 1.*

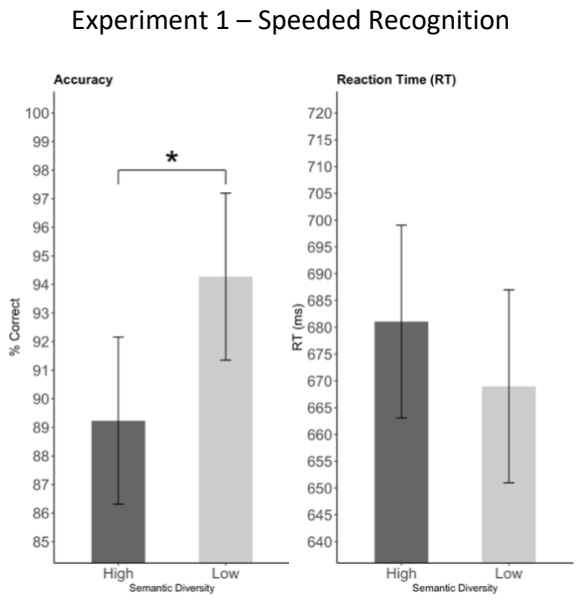
	Estimate	Standard Error	t value	p value
(a) Speeded recognition				
Intercept	-155.9	2.75	-56.77	<.001*
Sem. Diversity	0.62	0.87	0.71	.477
(b) Semantic relatedness				
Intercept	721.1	2.30	313.33	<.001*
Sem. Diversity	1.91	0.81	2.36	.019*

Note. To improve interpretability, estimate and standard error were multiplied by 100,000 in speeded recognition and by 100 in semantic relatedness.

Speeded recognition. Data from two participants were lost due to equipment failure, meaning these analyses are based on 38 participants. Only "yes" trials were analysed. Mean accuracy rates in the two diversity conditions are summarised in the left panel of Figure 4. Accuracy was significantly higher in the low-diversity condition (M = 94.4%) than in the high-diversity condition (M = 89.3%; z = 2.12, p = .034; see Table 2 for model summary). Turning to RT, only correct responses were analysed (91.8% of all "yes" trials). One trial was removed as it was more than 2.5 SD below the mean RT. Pseudowords experienced in the low-diversity condition elicited faster response (right panel of Figure 4), but this difference was small at 10 ms, and not statistically significant (t = -0.712, p = .477; see Table 3 for model summary).

Figure 4

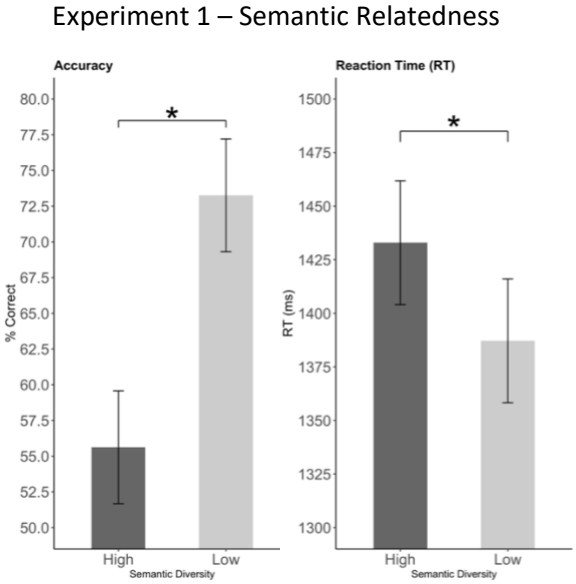
Mean accuracy (left) and RTs (right) in the two diversity conditions in speeded recognition, Experiment 1



Semantic relatedness judgement. Overall accuracy was above chance (*M* = 64.5%, *SD* = 11.4%), indicating that participants had gained some knowledge of the meanings of the novel pseudowords [one-sample t-test: $t(39) = 8.01$, $p < .001$]. The accuracy data are summarised by diversity condition in Figure 5, and model output is shown in Table 2. Performance was significantly more accurate in the low-diversity condition, $z = 6.60$, $p < .001$. RT to correct trials (14.2% error trials were excluded) patterned in the same direction with faster judgement to novel words encountered in the low-diversity passages ($t = 2.36$, $p = .019$; see Table 3 for model summary).

Figure 5

Mean accuracy (left) and RTs (right) in semantic relatedness judgement for words experienced in the two diversity conditions, Experiment 1.



Note. Error bars represent 95% within-subject confidence intervals.

Discussion

We induced variation in a novel word's semantic diversity at exposure and investigated whether this influenced the learning and processing of the novel words. Young adults read passages in which pseudowords (e.g., *ossanic*) replaced low-frequency words (e.g., *mendacious*). For each participant, five novel words appeared in passages built around the same discourse topic (low semantic diversity) while another five novel words appeared in passages spanning six distinct discourse topics (high semantic diversity). Word learning was assessed using three different measures, namely pseudoword inferiority, speeded recognition, and semantic relatedness. Across all three tasks, novel words experienced in the low-diversity condition were advantaged relative to the same words seen in the more diverse condition. Specifically, the finding that low-diversity novel words were recognized more accurately in speeded recognition contrasts with Johns et al. (2016),

who reported more accurate recognition for novel words experienced in *more* diverse contexts. Given the experimental approach is so similar, why should the results pattern in an opposite direction?

One possibility is that the reading materials in our experiment centred around familiar themes whereas Johns et al. (2016) used more obscure material. Consistent with this, our passages were rated as more comprehensible than those used by Johns et al. (see Figure 2). It is well-established that when a memory target—be it a novel word or a novel image—is encountered, it is encoded with reference to what is already known (e.g., Nelson & Shiffrin 2013). Novel items that are congruent with pre-existing knowledge are more likely to be remembered than those low in congruence (e.g., Greve, Cooper, Tibon, & Henson, 2019; Kaefer, Neuman, & Pinkham, 2015; Kan, Alexander, & Verfaellie, 2009). For word learning specifically, Pulido (2004) asked adult L2 learners of Spanish to read four Spanish texts, two describing familiar situations and two describing less familiar scenarios. Target words encountered in texts describing the more familiar scenarios were better retained, both immediately and 28 days later (see also Diakidoy, 1998; Nagy, Herman, & Anderson, 1985). Overall, there is converging evidence that incidental word learning, at least in its initial stages, fares better when the discourse topic is more familiar.

There are multiple explanations as to why greater topic familiarity is beneficial. For one, familiarity with the topic might help people infer the meaning of new words more easily (Freebody & Anderson, 1983). For another, when background knowledge is in place, comprehension resources can be allocated to processing the novel words, facilitating word learning (e.g., Cervetti, Wright, & Hwang, 2016; Chiesi, Spilich, & Voss, 1979, Horst, 2013). A different type of explanation builds from Steyvers and Tenenbaum (2005). When new words are introduced into the mental lexicon, they build associations with pre-existing words and knowledge. When topic familiarity is high, new words have greater opportunity to "anchor" to the topic and its pre-existing representations in long-term memory (Mak, 2019). Secure anchorage with long-term memory might help word learning and retention by reducing interference (e.g., Wang et al., 2017). In contrast, when the topic is less familiar, novel words are less likely to "anchor", resulting in more fragile memory representations of the words.

Topic familiarity was high across both diversity conditions in our experiment, yet more learning was seen for words experienced in the low-diversity condition. Arguably, the two exposure conditions nevertheless differed in the opportunity they offered for new words to anchor onto pre-existing knowledge. On this view, the low-diversity passages (i.e., six encounters in the same discourse topic) provided opportunity for more secure anchorage than the high-diversity passages (i.e., six encounters in six different discourse contexts). Future work is needed to examine what kind

of prior knowledge facilitates anchoring (e.g., Mak & Twitchell, 2020) and how this anchoring process may be influenced by the amount of prior knowledge accumulated over the life span. For instance, Ramscar, Sun, Hendrix, and Baayen (2017) reported that in older adults, accumulated associations that an existing word has impeded the formation of new associations, suggesting the possibility that older adults may not benefit from anchoring as much as young adults.

Returning to the question why our results differed to those obtained by Johns et al. (2016), high semantic diversity, as Johns et al. argued, may boost learning by encouraging readers to allocate more attentional resources to encoding the novel words. This is supported by their observation that reading times were significantly longer for high-diversity passages. In Experiment 1, we also observed the same reading times difference; despite this, a low-diversity advantage emerged in the post-tests, suggesting that the attention-related benefit associated with high-diversity environments was offset perhaps by topic familiarity, hence bringing about the different pattern of results.

A key limitation of Experiment 1 is that it is unclear the extent to which the observed low-diversity advantage can be attributed to a third factor such as arousal and differences in pre-existing knowledge and attitudes. For instance, the pseudoword *ossanic*, based on *mendacious*, appeared in six different passages centred on Donald Trump in the low-diversity condition; in the high-diversity condition, *ossanic* was embedded in six different-themed passages: one about North Korea, one about Nigel Farage, and one about childhood memory, etc. It is reasonable to argue that the six Donald Trump passages were more similar in terms of their level of arousal and the amount of background knowledge required. In other words, the low-diversity advantage in Experiment 1 can be attributed, at least partially, to these differences instead of semantic diversity *per se*. In light of this, we conducted a computational study where we simulated spreading activation in lexical networks that differed by semantic diversity.

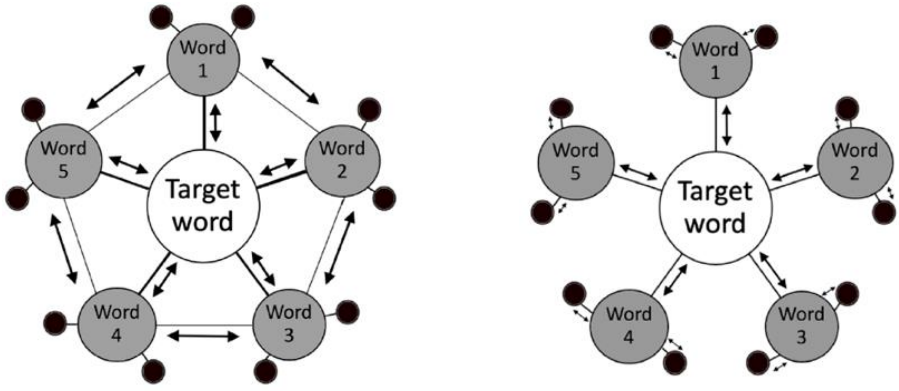
Simulation 1

We used the reading materials developed for Experiment 1 to investigate whether target words appearing in the low-diversity passage would show more efficient processing as estimated by a computational simulation. This offered an opportunity to replicate the behavioural findings, and to see whether contextual variation influences how words are learned and processed when other factors such as arousal become irrelevant. Before going into the details of the simulation, we first need to consider the background from which the simulations were derived.

According to localist approaches, a word is represented in the mental lexicon by a distinct node in a network, and words are connected if they are phonologically, orthographically, or semantically related (e.g., Collins & Loftus, 1975). An accumulating body of evidence suggests that where the word stands in a network influences how efficiently its form is processed (e.g., Chan & Vitevitch, 2009, 2010; Siew & Vitevitch, 2016; Mak, Hsiao & Nation, 2020). Of particular interest here is a network structure known as clustering coefficient. This refers to the degree to which neighbours of a word node are also neighbours of each other. In a word learning experiment, Goldstein and Vitevitch (2014) found that adult participants retained more novel words with higher clustering coefficients. They argued that for such words, the greater interconnections among the immediate neighbours (see left panel of Figure 6) help activations circulate back to the novel word node (note: activation is assumed to spread bilaterally), thereby strengthening its nascent representation. In contrast, for novel words with a lower clustering coefficient (see right panel of Figure 6), activation is more likely to disperse outside of the immediate neighbourhood, reducing the amount of activation circulating back to the novel word target. In turn, this lowers its likelihood of being retained.

Figure 6

Left: *A target word with a higher clustering coefficient.* ***Right:*** *A target word with a lower clustering coefficient.*



Given Goldstein and Vitevitch's (2014) finding, we hypothesised that the low semantic diversity condition in Experiment 1 led to higher clustering coefficients for the novel words (relative to those in the high diversity condition) thereby boosting the amount of activations they received and leading to better learning. To test this idea, we first computed term-term co-occurrence matrices for each target word, using the reading materials from Experiment 1. Two sets of networks were then constructed for each target word, one low-diversity and one high-diversity. Here, the networks are assumed to reflect the target word's contextual history, which in turn exerts an influence on the efficiency with which the word is identified in tasks such as speeded recognition. Next, spreading activation was simulated across the networks, allowing us to see if target words appearing in the low-diversity network receives more activation. We used "spreadr", a package that has simulated behavioural data from psycholinguistic tasks that tap single-word recognition (Siew, 2019). It is unclear whether the package can be extended to more complex tasks such as semantic judgement. Therefore, our simulation data are most appropriately considered against the behavioural data from the speeded recognition task in Experiment 1.

Like all computational simulations, some assumptions are necessary. Concerning topic familiarity, we assumed that if a target word is experienced in familiar topics, it will be connected to well-established nodes that send activations back, thereby reinforcing the target's representation. In contrast, if a target word is experienced in unfamiliar topics, it is assumed that the target will be connected to impoverished or previously non-existent nodes, resulting in less activation being sent back. In the present simulation, activations in all the networks were allowed to flow relatively freely; this is to reflect that all the passages in Experiment 1 were based on familiar topics, which could presumably reinforce the targets by sending back activations. If one wished to simulate the effect of *unfamiliar* topics, the free flow of activations could be restricted so that the target node receives fewer supportive activations from its neighbouring nodes. In addition, activations in both the high- and low-diversity networks were allowed to spread at the same rate, meaning that every topic is set to be "equally capable" at sending activations back to the target. This way, we were able to eliminate the potential influence of arousal and required background knowledge.

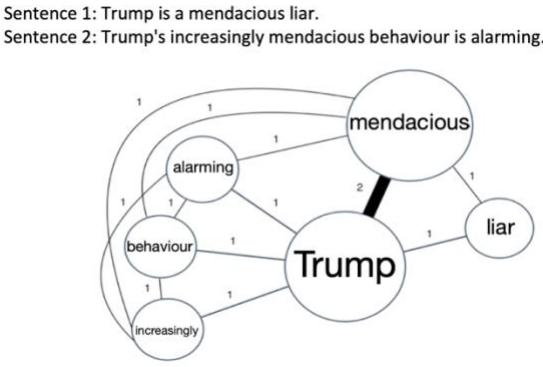
Method

We used eight of the 10 base words from Experiment 1 for this simulation (*mendacious*, *predilection*, *constellation*, *proliferation*, *diffident*, *latent, vagarious,* and *castigate*). [1] Two corpora were created for each word, one for the low-diversity passages and one for the high-diversity passages, totalling eight for each diversity condition. Punctuation, numerals, and function words (e.g., prepositions, pronouns) were removed. Proper nouns containing two words (e.g., New York) were merged (e.g., NewYork), and the corpora were stemmed using the "SnowballC" package (Bouchet-Valat, 2019) in R. Next, a global term-term co-occurrence matrix was created for each corpus, using the "chinese.misc" package (Wu, 2019) on the basis that co-occurrence provides "an objective measure that aggregates semantic, conceptual and syntactic relatedness" (Beckage, Smith, & Hills, 2011, p. 2). Unlike some previous studies which captured co-occurrence within a 5 to 10 word window around a target word (e.g., Beckage et al., 2011; Hills, Maouene, Riordan, Smith, 2010; Li, Burgess, & Lund, 2000; McDonald & Ramscar, 2001), we opted for global co-occurrence computed across passages. Our reasoning was threefold. First, corpus size for each word was small (~400 words). Second, global co-occurrence provided an opportunity to capture the topic of each passage. For example, with a window size smaller than 10, the target word *mendacious* would not have "co-occurred" with *Trump* in most of the passages. Third, the passage was the "unit" used in Experiment 1, and was therefore retained here, rather than determining an arbitrary fixed window.

Sixteen co-occurrence matrices were created and transformed into a weighted, undirected network using the "igraph" package (Csardi & Nepusz, 2006). Figure 7 depicts a sample network created for two illustrative sentences.

Figure 7

A sample weighted undirected network.



Note. The number next to each link represents how many times two words co-occurred together in the corpus. This is also the *weight* of connection.

[1] Two of the 10 target words (i.e., *avidity*, *shenanigans*) in Experiment 1 shared the same contextual environments with two other target words (i.e., *diffident*, *mendacious*). To avoid duplications, *avidity* and *shenanigans* were excluded.

Spreading activation was then implemented in each network using the "spreadr" package (Siew, 2019). There are several parameters that can be specified in the simulation; full details of settings are available in Appendix D. Of note here is retention. This refers to the proportion of activation that remains in a node at each time step. For example, a value of 0.8 means that at each time step, 80% of activation that a node receives will be retained in the node while the remaining 20% will spread to the neighbouring nodes. Following Siew (2019) and Vitevitch et al. (2011), we used nine retention values ranging from 0.1 to 0.9, in increments of 0.1. This was to ensure that the outcome of the simulations is generalizable across retention levels (Siew, 2019). Spreading activation within a network produced an activation value for each node at every time step. Following Siew (2019, p. 916), "the activation value of the target node at the final time step was taken as a proxy for the efficiency with which the word was retrieved from the lexicon. Specifically, higher activations [at the final time step] correspond to faster RTs and/or higher accuracy rates in psycholinguistic tasks". To summarise, we ran 144 simulations (8 words x 2 diversity conditions x 9 retention levels), 72 as high semantic diversity networks and 72 as low diversity networks.

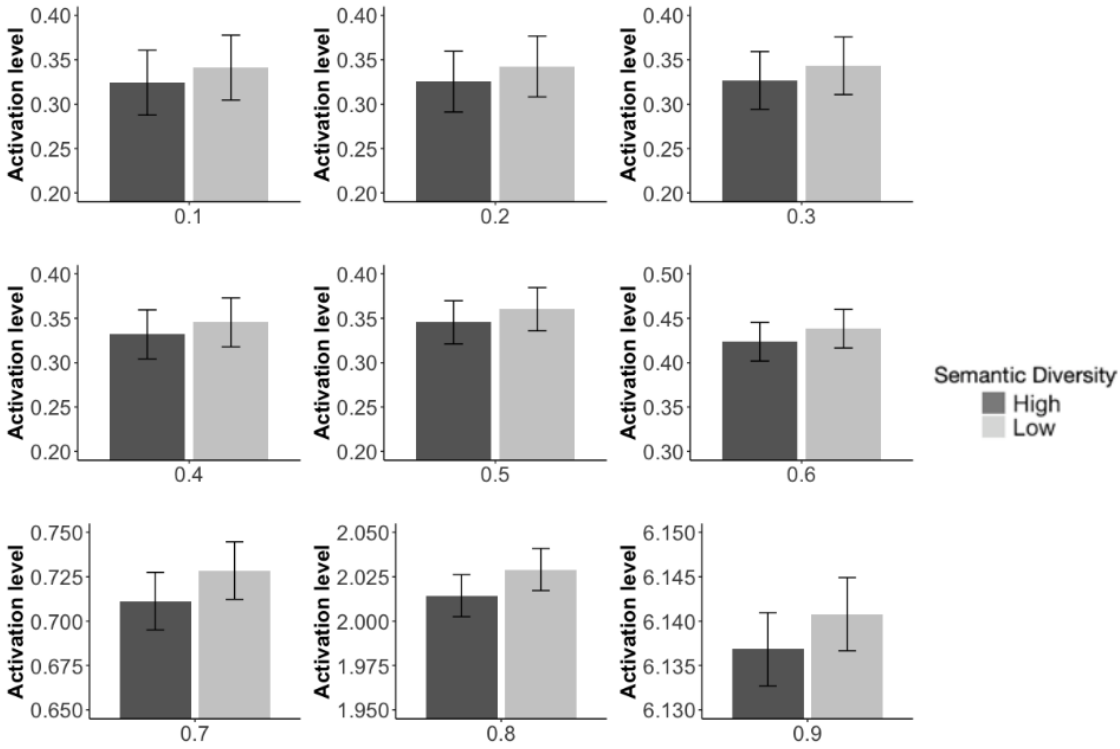
Simulation results

Figure 8 shows the average activation levels in the two types of diversity networks across the nine retention levels. Mean activation of the target words was consistently higher in the low-diversity networks than in the high-diversity networks across all the retention levels. To test whether the difference is statistically significant, we used a Wilcoxon signed-rank test, as the distribution was non-normal. It confirmed that activation levels of the target words were reliably higher in the low-diversity networks (M = 1.230) than in the high-diversity networks (M = 1.216), $Z(143) = 2.66$, $p = .004$.

Next, we obtained the clustering coefficients of the target nodes in the two types of diversity networks. As shown in Figure 9, clustering coefficients of the target nodes were consistently higher in the low-diversity networks (M_{High} = 0.206 vs. M_{Low} = 0.249), $t(7) = -6.06$, $p < .001$.

Figure 8

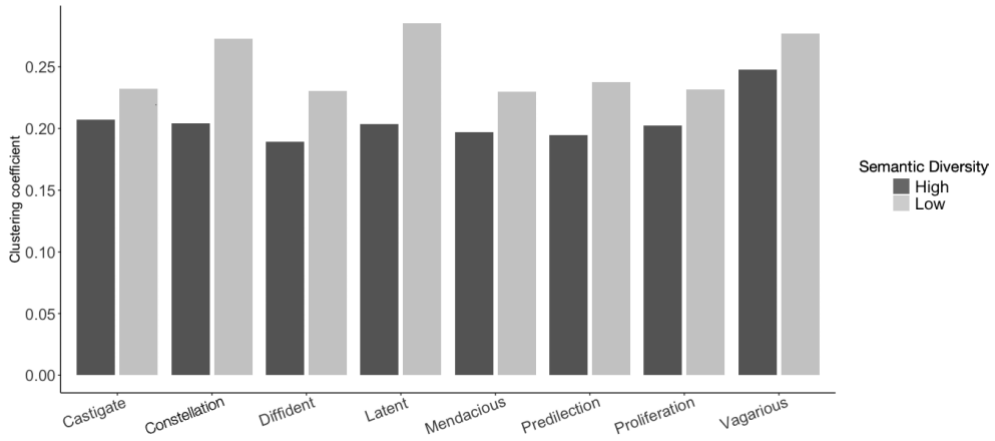
Mean final activation values in the target node in each type of diversity networks across the nine retention levels in Simulation 1.



Note. Error bars represent 95% within-item confidence intervals.

Figure 9

Clustering coefficients of the target base words in the two types of diversity networks in Simulation 1.



Discussion

Simulation 1 provided computational evidence that in the initial stages of word learning, words experienced in low-diversity (but topically familiar) contexts are advantaged in visual recognition relative to the same number of encounters in more diverse contexts. This processing

advantage aligns with the behavioural data in Experiment 1, where words experienced in the low-diversity condition were better recognized. Simulation 1 eliminated some of the potentially confounding variables from Experiment 1, including for example arousal and the different levels of required background knowledge. The emergence of a low-diversity advantage in the simulation data therefore supports our behavioural finding that low-diversity contexts are more conducive to the initial stages of incidental word learning.

The target nodes received more activation when they appeared in the low-diversity networks, plausibly because these nodes had higher clustering coefficients, reflecting a more dense immediate neighbourhood. This interpretation stems from Goldstein and Vitevitch (2014) who found that novel word forms with higher clustering coefficients were better recalled in their word learning study. One benefit of having a high clustering coefficient, according to the authors, is that it helps activation to circulate back to the target nodes, which in turn strengthens their nascent representations. In short, situating new words in more close-knit and familiar contexts may boost initial learning. This description also captures the notion of anchoring as discussed in Experiment 1 in that low-diversity passages allowed the novel words to be secured onto long-term knowledge, thereby boosting word learning and retention.

Finally, we stress that the current simulations have at least two limitations. First, they were built upon assumptions derived from spreading activation dynamics (Anderson, 1983; Collins & Loftus, 1975). For instance, it was assumed that a word is represented as a single node in the lexicon. While our simulations serve as an existence proof of the behavioural findings, they do not rule out other possible accounts based on other assumptions. For example, it is possible that our behavioural findings could fit with a connectionist framework (e.g., Harm & Seidenberg, 2004; McClelland & Rumelhart, 1981; Plunkett, 1998), where a word is not represented as a single node but as distributed activity patterns. Future work is needed to model our behavioural findings within other theoretical and computational frameworks. Second, the networks we constructed were based solely on the reading materials we developed; for instance, the passages with *mendacious* as the target word also included words like *Trump*, *tweets*, *President,* and *Republican* and these words were also present in the networks. These words represent (presumably) familiar concepts, meaning that in the simulation, they were set to be capable of reinforcing the target node by sending back activations. In other words, the networks had some prior knowledge (e.g., *Trump*, *President*, *Republican*) built into them. Clearly, however, the amount is limited and is some distance from human learning, where word learning is influenced by a large and complex pre-existing knowledge base (e.g., Ramscar et al., 2014; 2017).

In summary, Experiment 1 and Simulation 1 both suggested that the initial stages of word learning can profit from words being experienced in familiar passages that are topically similar, relative to the same amount of experience in familiar passages that are contextually diverse. Importantly however, this low-diversity learning advantage is at odds with what is known about established words where there is ample existing evidence to show that high semantic diversity words are identified more efficiently, even when frequency and other key lexical statistics are controlled for (e.g., Hoffman & Woollams, 2015; Hsiao et al., 2019; Jones et al., 2012; Pagán et al., 2020). The contrasting results between these effects on established words and our findings for newly learned words suggests that there is a switch in direction in how contextual variation influences visual word recognition. With this as a backdrop, we explored in Simulation 2 how a low-diversity advantage in the initial stage of word learning becomes a high-diversity advantage as word knowledge matures.

Simulation 2

Data from Experiment 1 and Simulation 1 converged to show a processing advantage for novel words experienced in contexts that are less semantically diverse. However, given there is clear evidence that known words higher on semantic diversity are identified more accurately and rapidly in lexical decision (e.g., Hoffman & Woollams, 2015; Hsiao & Nation, 2018), we explored in Simulation 2 how does a high-diversity advantage might come to be.

We hypothesized that if new words are securely anchored to a familiar topic, word learning might be better placed to take advantage of contextual variation. In Simulation 2, we used the same simulation method as before, but all target words were repeatedly embedded in one discourse topic before contextual variation was induced at a second step. This design is best explained with reference to Table 4.

Table 4

How semantic diversity was implemented in Simulation 2 (and Experiment 2).

	Anchoring passages					**Post-Anchoring passages**		
	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5	Passage 6 (each contains 2 target words)	Passage 7 (each contains 2 target words)	Passage 8 (each contains 2 target words)
High-diversity word	Topic 1	Topic 1	Topic 1	Topic 1	Topic 1	Topic 2	Topic 2	Topic 2
Low-diversity word	Topic 2	Topic 2	Topic 2	Topic 2	Topic 2			

Passages 1 to 5 constituted the anchoring phase. Here, all words, regardless of the diversity condition they were assigned to, appeared in passages centred around one discourse topic. After anchoring, the diversity manipulation was introduced. Target words in the high-diversity condition appeared in a different topic in passages 6 to 8. In contrast, target words in the low-diversity condition remained within the same topic throughout all eight passages. With some anchorage in place, we predicted that target words in the high-diversity condition would show a processing advantage.

Note that, the number of exposure per word was eight here, not six as in Experiment/Simulation 1. We purposely did not match the number of exposures as we hypothesized that a high-diversity advantage would emerge in Simulation 2 when the novel words were securely anchored. Experiment/Simulation 1 provided evidence of secure anchorage following six exposures. To test our hypothesis, therefore, it was necessary to increase the number of exposure beyond six.

Materials

The materials for Simulation 2 were designed to be used in a subsequent behavioural experiment, reported later in this paper. Therefore, the passages were constructed to be appropriate for young adults to read in approximately 45 minutes, allowing us to maintain the single exposure session used in Experiment 1. All passages had high topic familiarity.

We used the same 10 base words as in Experiment 1 (e.g., *mendacious*) plus two additional words (e.g., *architecture*, *imbroglio*) to increase power. As before, two sets of passages were created for each target word: The low-diversity set comprised eight passages; their discourse topic was the same (e.g., Donald Trump). The high-diversity set comprised five passages on one topic (e.g., a Michael Jackson documentary) and three passages on another topic (e.g., Donald Trump).

Each target word, regardless of its assigned diversity condition, appeared in one discourse topic for passages 1 to 5 (see Table 4). These passages are referred to as the "anchoring passages". A total of 120 anchoring passages were created (= 12 target words x 5 exposure x 2 diversity), each containing one target word and an average of 36.2 words. The high and low diversity conditions were matched on the number of words (M_{High} = 37.3 vs. M_{Low} = 35.1; $t(119) = 1.62$, $p = .11$) and readability, as indexed by Flesch Reading Ease test (M_{High} = 41.2 vs. M_{Low}= 45.1; $t(119) = -0.71$, $p = .49$).

In addition, 18 passages from Experiment 1 were chosen and modified to serve as the post-anchoring passages (passages 6 to 8, see Table 4). Each contained two target words, one belonging to the high-diversity condition and one to the low-diversity condition.

Simulation procedure

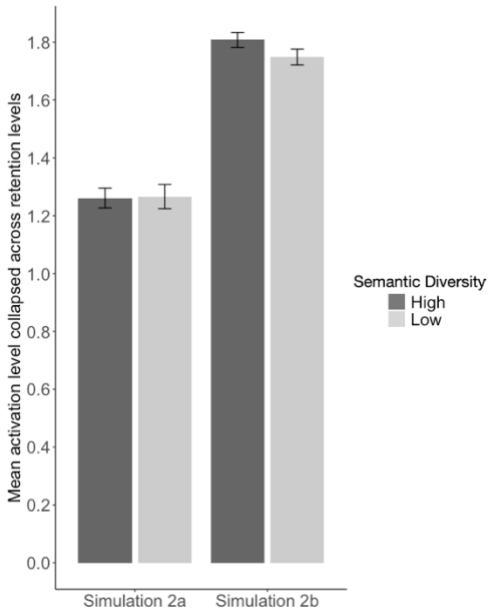
Spreading activation was simulated in broadly the same way as Simulation 1, using the same procedures and packages. Simulation 2a used only the anchoring passages as input (i.e. passages 1-5). Since the anchoring passages were "identical" across the two diversity conditions, no difference in activation values was expected. Simulation 2b included the post-anchoring passages as well as the anchoring passages. If post-anchoring semantic diversity influences the structure of the networks, there should be a difference in activation levels between the two conditions in Simulation 2b. Of interest is how this pattern compared to Simulation 1 which did not contain an anchoring phase. We hypothesized that adding diversity post-anchoring should result in the target words in the high-diversity networks receiving more activation than those in the low-diversity networks, in contrast to Simulation 1. Overall, we ran a total of 432 simulations, half for Simulation 2a, the other half for Simulation 2b. For each set of simulation, 108 were on high-diversity networks and 108 on low-diversity networks, each comprising 12 words x 9 retention levels.

Results

Figure 10 shows the mean activation values of the target nodes, collapsed across retention levels, between the high and low diversity networks in both simulations. Simulation 2a included only the anchoring passages. As expected, and as shown in Figure 10, there was no difference in activation values between target nodes appearing in the high (*M* = 1.261) and in the low (*M* = 1.265) diversity networks. This non-difference was confirmed in a Wilcoxon signed-rank test, *Z*(107) = -0.55, *p* = .292.

Figure 10

Mean activation values of the target nodes, collapsed across retention levels, between the high and low diversity networks in Simulations 2a and 2b.

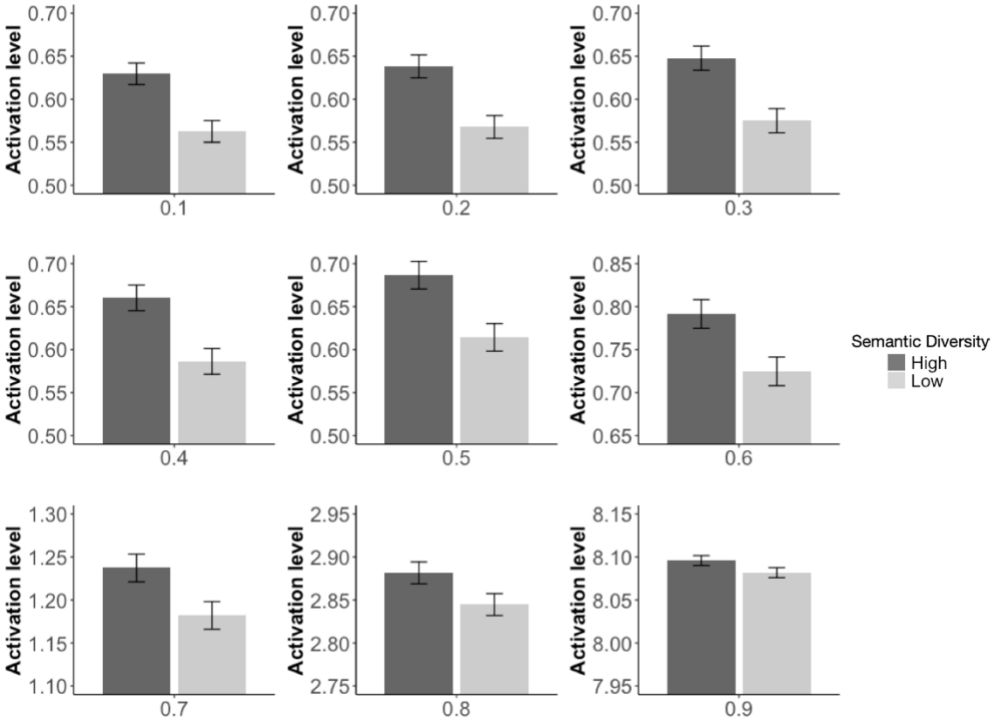


Note. Error bars represent 95% within-item confidence intervals.

In contrast, Simulation 2b, which included the post-anchoring passages into the networks, produced higher mean activation levels in the high-diversity networks (*M* = 1.807) than in the low-diversity networks (*M* = 1.748; see Figure 10). A Wilcoxon signed-rank test confirmed that target nodes in the high-diversity networks received significantly more activations than those in the low-diversity networks, *Z*(107) = -8.90, *p* < .001. This difference was also evident across the nine retention levels, shown in Figure 11.

Figure 11

Mean final activation values in the target node in each type of diversity networks across the nine retention levels in Simulation 2b.

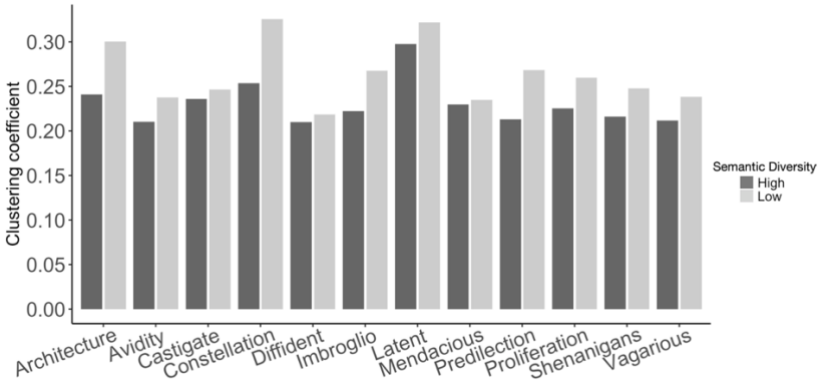


Note. Error bars represent 95% within-item confidence intervals.

Finally, we checked the clustering coefficients of the target nodes (i.e., the probability with which the neighbours of the target are also neighbours of each other) in the two types of diversity networks. As in Simulation 1, clustering coefficients of the target nodes were consistently higher in the low-diversity condition (M_{High} = 0.231 vs. M_{Low} = 0.264; see Figure 12), $t(11) = -3.30$, $p < .001$.

Figure 12

Clustering coefficients of the target base words in the two types of diversity networks in Simulation 2b.



Discussion

Simulation 2 tested the hypothesis that if new words are securely anchored to a familiar topic, word learning might be better placed to take advantage of contextual variation. In Simulation 2a, there was no difference in the amount of activation for target words in the two types of diversity networks, as expected. However, in Simulation 2b, adding diversity via the post-anchoring passages changed the pattern of results such that target words in the high-diversity networks received reliably more activations. Previous work has attributed activation level to efficiency of lexical processing (e.g., Siew, 2019). The results of Simulation 2b are therefore consistent with corpus-based studies in showing a recognition advantage for words higher in semantic diversity (e.g., Hoffmann & Woollams, 2015; Hsiao & Nation, 2018; Jones et al., 2012).

We next considered why the novel words appearing in the high-diversity networks received more activations in Simulation 2b. In Simulation 1, the low-diversity advantage was attributed to higher clustering coefficients. Clearly, this cannot explain the high-diversity advantage seen in Simulation 2b, as the clustering coefficients were also always lower in the high-diversity condition (M_{High} = 0.231 vs. M_{Low} = 0.264). Instead, increased activation in the high-diversity condition may stem from the target words being associated with one additional discourse topic, meaning that these targets have connections with more unique word nodes (M_{High} = 203.3 vs. M_{Low} = 188.8; $t(11)$ = 6.05, p = < .001). Having connections with more unique word nodes allows more activations to circulate back to the target node, thereby reinforcing its activation. Nevertheless, this was also the case in Simulation 1, where the number of unique word nodes was also higher in high-diversity networks (M_{High} = 276.6 vs. M_{Low} = 237.1; $t(7)$ = 5.28, p = .001). Given Simulation 1 did not find a high-diversity advantage, it suggests that being connected with more unique word nodes alone cannot account for the high-diversity advantage in Simulation 2b. One explanation is that in addition to being connected to more unique word nodes, the target words in the high-diversity networks in Simulation 2b also had a relatively close-knit immediate neighbourhood, established as a result of the anchoring passages. This appears to be the case given the mean clustering coefficients of the high-diversity words increased from 0.206 in Simulation 1 to 0.231 in Simulation 2b. Overall, it is reasonable to attribute the high-diversity advantage found in Simulation 2b to two factors: (i) an anchoring phase that produced more close-knit neighbourhoods compared to those in Simulation 1, and (ii) an opportunity for building associations with more unique word nodes, which further helps activations to circulate back to the targets.

Finally, it is worth comparing the low-diversity conditions in Simulations 1 and 2b. In Simulation 1, the target words in the low-diversity condition were anchored onto one topic six times while in Simulation 2b, they were anchored onto one topic eight times. If anchoring was always

beneficial to word learning, we might expect the words in the low-diversity networks of Simulation 2b to receive more activations than those in the high-diversity networks. However, the opposite was true in Simulation 2b. In light of this, we suggest that the effect of anchoring may *saturate* (e.g., Murray & Forster, 2004); that is, repeated anchoring beyond a certain threshold might have little or even no detectable effect on subsequent processing. In turn, this suggests that while anchoring (i.e., experiencing new words in similar and familiar contexts) may be conducive to the initial stage of word learning, additional anchoring beyond a certain threshold may lead to little change in learning outcomes.

Finally, caution is required when interpreting the results of Simulation 2b: This kind of network simulation does not take into account the effect of learning order, meaning that Simulation 2b would have produced the same result had the order of topics in the high-diversity condition been intermixed (i.e., Topic 11111222 and Topic 12211211 would produce identical simulation results). Therefore, while Simulation 2b provides evidence for anchoring being a potential contributor to a high-diversity advantage in word learning, it does not speak to whether anchoring must precede diversity. In line with our account, however, a computational study by Jorge-Botana, Olmos, and Sanjosé (2017) suggested that anchoring preceding diversity might be the most conducive growth path for word knowledge. Future experimental work is needed to determine the importance of learning order.

To sum up, Simulation 2 sets up the prediction that participants should show a learning advantage for words experienced in diverse contexts, if the words were given an opportunity to anchor onto one discourse topic. This prediction was tested in Experiment 2.

Experiment 2

As in Experiment 1, young adults read passages containing novel words in high vs. low semantic diversity contexts. The reading materials were taken from Simulation 2. Unlike Experiment 1, the target words in the high-diversity conditions were given an anchoring opportunity (i.e., repeatedly experiencing a word in a familiar topic) in the first five passages before diversity was introduced, following the procedure developed for Simulation 2 (see Table 4). In light of the results from Simulation 2, our hypothesis was that, given an anchoring opportunity, target words experienced in the high-diversity condition would show a processing advantage in visual recognition. Specifically, it was predicted that they would be recognized more efficiently in speeded recognition (greater accuracy and/or faster RTs) and that they would induce greater top-down interference in the pseudoword inferiority task, leading to poorer performance. As for semantic judgement, we predicted that high-diversity novel words would be judged more poorly than the low-diversity words, in line with Experiment 1 and Johns et al. (2016). Experiment 2 was pre-registered ahead of data collection (https://aspredicted.org/blind.php?x=ti8g6f).[2]

Method

Participants

Forty-three undergraduate psychology students from Oxford University (11 males; M_{age} = 19.4; SD_{age} = 1.49) participated for course credits. All were native speakers of English with normal or corrected-to-normal vision and all reported no known history of any developmental disorders. None of them had taken part in Experiment 1. Following the pre-registered exclusionary criteria, three participants were excluded: two because they scored 2 SDs below the sample mean on the comprehension questions, and one because their accuracy rates in the speeded recognition task fell below 75%. All analyses, unless otherwise specified, were based on the remaining 40 participants.

Materials and Procedure

We used the reading materials developed for Simulation 2. As in Experiment 1, the 12 target base words (e.g., *mendacious*) were replaced by pronounceable pseudowords (e.g., *mendacious* → ossanic). The general procedure also followed Experiment 1. Participants first completed an exposure phase where they read the passages; they then completed a surprise test phase designed to measure target word learning.

[2] The pre-registration document included an additional post-test, namely importance judgement. As this did not feature in Experiment 1, we chose not to report it here to avoid over-complication. Further information can be obtained from the first author.

As per Experiment 1, semantic diversity was manipulated within-subject. Each participant saw a total of 12 novel words, six in the high-diversity condition and six in the low-diversity condition. Assignment to the diversity conditions was fully counterbalanced across items. Participants saw each item eight times, five times in the anchoring phase and three times in the post-anchoring phase. The diversity manipulation was induced in the post-anchoring phase; prior to this, the experience of high vs. low diversity words was identical in the anchoring phase. Note that the high-diversity conditions in Experiment 1 and 2, despite sharing the same name, were somewhat different from each other (i.e., six topics in Experiment 1, but only two topics in Experiment 2).

(i) Exposure phase

The exposure phase in Experiment 2 had two components: (i) anchoring passages and (ii) post-anchoring passages (see Table 4). Each participant read a total of 60 anchoring passages and then 18 post-anchoring passages. For each anchoring passage, participants answered one yes/no question. Pilot testing confirmed that question difficulty was equivalent across the diversity conditions. Each passage contained one exposure to one target word. For the post-anchoring passages, participants rated the comprehensibility of each passage and answered two comprehension questions (taken from Experiment 1). Each of these passage contained two target words, one belonging to the high-diversity condition, another to the low-diversity condition. All participants read the same 18 post-anchoring passages.

(ii) Test phase

Learning was assessed with three post-tests, namely pseudoword inferiority, speeded recognition, and semantic relatedness judgement. Task parameters and procedures were identical to Experiment 1.

Results

The analytic approach followed the pre-registered analytic plan and was identical to that described in Experiment 1.

(i) Exposure phase

On average, participants achieved 95.0% (*SD* = 2.13%) and 82.2% (*SD* = 2.93%) accuracy in the comprehension questions associated with the anchoring and post-anchoring passages, respectively. There was no accuracy difference between the high- and low-diversity passages in the anchoring phase, $t(39) = 0.28$, $p = .89$. In the post-anchoring phase, since each passage contained

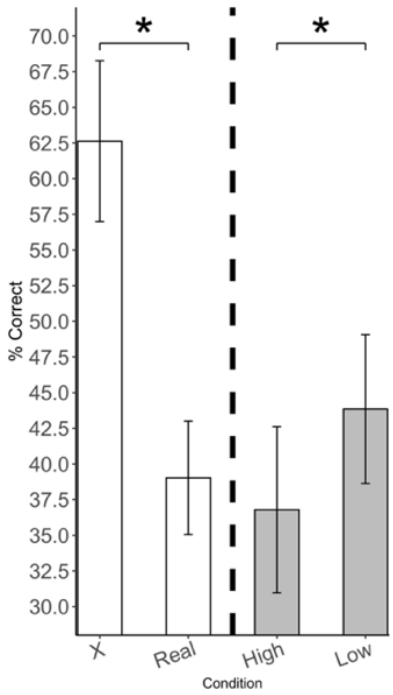
one high-diversity and one low-diversity novel words, it was not possible to compare the accuracy rates across the two conditions.

(ii) Test phase

Pseudoword Inferiority. Mean accuracy rates of the four conditions are summarised in Figure 13. Letter identification in the real word condition (*M* = 37.9%) was significantly poorer than in the X condition (*M* = 60.7%) (*z* = 3.26, *p* = .001), replicating the pseudoword inferiority effect reported by Grainger and Jacobs (2005) and observed in Experiment 1. Letter identification was also significantly poorer in both diversity conditions than the X condition (*z*s > -4, *p*s < .001). This indicates that the novel words in both conditions had been learned sufficiently to bias letter identification. Finally, there was also a significant difference between these two diversity conditions, with novel words in the high-diversity passages (*M* = 35.7%) biasing letter identification to a greater extent than those in the low-diversity passages (*M* = 42.5%; *z* = -2.42, *p* = .016; see Table 5).

Figure 13

Mean accuracy rates across the four conditions in the pseudoword inferiority task, Experiment 2.



Note. (i) Error bars represent 95% within-subject confidence intervals. (ii) Lower accuracy in the real and high-diversity conditions is indicative of greater top-down interference.

Table 5

*Model summary of the GLME models examining the effect of semantic diversity on **accuracy** in each of the post-tests, Experiment 2.*

	Estimate	Standard Error	z value	p value
(a) Pseudoword inferiority				
Intercept	-0.52	0.21	-2.44	.015*
Sem Diversity	-0.17	0.07	-2.42	.016*
(b) Speeded recognition				
Intercept	3.51	0.30	11.64	<.001*
Sem Diversity	0.08	0.25	0.30	.762
(c) Semantic relatedness				
Intercept	1.24	0.16	7.70	<.001*
Sem Diversity	0.08	0.06	1.43	.152

Table 6

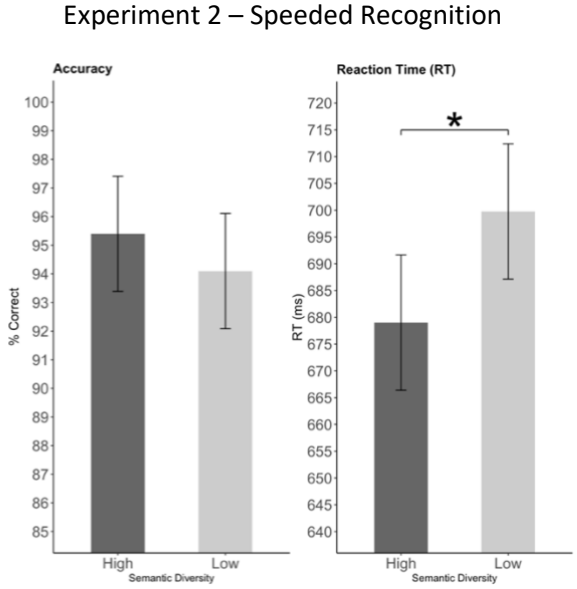
*Model summary of the LME models examining the effect of semantic diversity on **RTs** in each of the post-tests, Experiment 2.*

	Estimate	Standard Error	t value	p value
(a) Speeded recognition				
Intercept	-153.2	3.04	-50.34	<.001*
Sem Diversity	-2.21	0.75	-2.92	.004*
(b) Semantic relatedness				
Intercept	718.68	2.34	307.21	<.001*
Sem Diversity	-0.92	0.97	-0.94	.36

Note. To improve interpretability, estimate and standard error were multiplied by 100,000 in speeded recognition and by 100 in semantic relatedness.

Speeded recognition. Data from two participants were removed as they appeared to use the wrong response keys. Mean accuracy rates in the two diversity conditions are summarised in the left panel of Figure 14. Target words in the high-diversity condition (*M* = 95.4%) were recognised more accurately than those in the low-diversity condition (*M* = 94.1%), but this was not statistically significant (*z* = -0.30, *p* = .762; see Table 5 for model summary). Turning to RTs, mean RTs (to correct 'yes' trials) in the two diversity conditions are summarised in the right panel of Figure 14. Participants were significantly faster at recognizing words that were encountered in the high-diversity passages (*M* = 679 ms) than those in the low-diversity passages (*M* = 700 ms; *t* = -2.92, *p* = .004, see Table 6).

Figure 14
Mean accuracy (left) and RTs (right) in the two semantic diversity conditions in speeded recognition, Experiment 2.

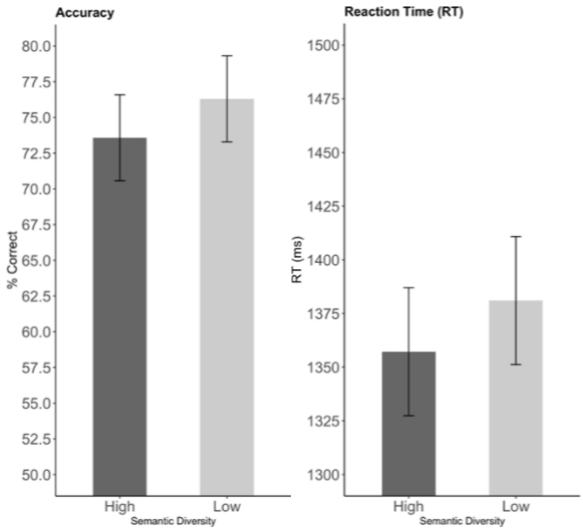


Note. Error bars represent 95% within-subject confidence intervals.

Semantic Relatedness. Overall accuracy was satisfactory (*M* = 75%, *SD* = 14.7%) and significantly above chance, $t(39) = 9.63$, $p < .001$. Mean accuracy rates in the two diversity conditions are plotted in the left panel of Figure 15. The accompanying GLME model, summarised in Table 5, found no reliable difference between the two diversity conditions ($z = 1.43$, $p = .152$). Mean RTs (to correct trials) in the two diversity conditions are plotted in the right panel of Figure 15. While target words in the high-diversity condition were judged more quickly, the difference was not significant ($t = -0.94$, $p = .36$; see Table 6 for model summary).

Figure 15
Mean accuracy (left) and RTs (right) in the two semantic diversity conditions in semantic relatedness judgement, Experiment 2.

Experiment 2 – Semantic Relatedness



Note. Error bars represent 95% within-subject confidence intervals.

Discussion

Experiment 2 examined the effect of semantic diversity on incidental word learning during independent reading. Young adults read passages containing pseudowords (e.g., OSSANIC) whose meanings were derived from words such as *mendacious.* Experiment 2 differed from Experiment 1 in one major respect: In Experiment 2, pseudowords in the high-diversity condition were first experienced in one familiar topic in the anchoring passages before diversity was introduced. Items experienced in this condition (vs. those in the low-diversity condition) showed a greater pseudoword inferiority effect and were recognized faster in speeded recognition. Although the mean RT advantage enjoyed by the high-diversity words in speeded recognition was small at 21 ms, it is comparable with the 26-ms advantage reported by Johns et al. (2016). These findings align with the data reported in Simulation 2, which showed that diversity, coupled with an anchoring opportunity, could boost the amount of activation received by the target words.

However, the findings here contrast with those from Experiment 1, which demonstrated a low-diversity advantage in word learning. We attribute the discrepancy to the fact that in Experiment 2, the high-diversity words were given an anchoring opportunity. Anchoring refers to the process by which new lexical knowledge is repeatedly experienced in one familiar topic such that it could be secured onto long-term memory. This might help stabilize and strengthen the representations of the novel words (Jorge-Botana et al., 2017; Mak, 2019), allowing them to be better placed to take advantage of contextual variation. We return to this in General Discussion.

While an anchoring account fits with the high-diversity advantage seen here, there are at least two other plausible accounts. First, high-diversity words in Experiment 1 were experienced in

six topics whereas in Experiment 2 they were only experienced in two topics. It is plausible that a little bit of diversity, as in Experiment 2, is conducive to word learning, but too much diversity, as in Experiment 1, is not. We cannot rule out this possibility, but this explanation would be at odds with Johns et al.'s (2016) finding that high-diversity novel words were advantaged in visual recognition, despite being experienced in five distinct topics. Second, high-diversity words in Experiment 1 were experienced once in each topic, but in Experiment 2, they were experienced five or three times in each topic, so perhaps high semantic diversity is only beneficial when readers have multiple opportunities to map the word to each topic. This is plausible, but again, it appears at odds with the high-diversity advantage reported by Johns et al. (2016), where the high-diversity words were only experienced once in each topic. The design discrepancy between Experiments 1 and 2 does not allow us to rule out these alternative explanations and more work is needed to test further whether anchoring opportunities before experiencing diversity drives a high-diversity advantage, as observed in Experiment 2. In the meantime, anchoring provides a ready explanation both for our findings and those described by Johns et al. (2016).

Moving onto the semantic relatedness judgement task: There was no effect of diversity on the learning of meaning in Experiment 2, as shown in both the accuracy and RT data. This contrasts with the findings of Experiment 1 where low-diversity words were easier to judge (see also Johns et al., 2016). This discrepancy may be due to the fact that in Experiment 2 words in the high-diversity condition were experienced in two topics only, compared with six topics in Experiment 1 (and five topics in Johns et al., 2016). Potentially, therefore, the meanings of the high-diversity words in Experiment 2 had not become complex or nuanced enough to hinder performance.

General Discussion

Our investigation was inspired by the finding that item-level contextual history, as captured by corpus-derived statistics such as semantic diversity, influences lexical processing: Known words that have been experienced in more varied contexts are easier to recognise in tasks such as lexical decision but more difficult to reflect on in tasks such as semantic judgement (e.g., Hoffman & Woollams, 2015; Hsiao et al., 2019; Jones et al., 2012). While consistent in showing an effect of contextual variation, research in this tradition cannot explain the nature of its effect, nor how it develops and accrues over time, as words are experienced. Instead, word-learning experiments are needed which manipulate contextual experience directly and track its effects on learning. We adopted this approach in two experiments. Novel pseudowords were experienced in either high or low diversity passages: High-diversity passages spanned multiple distinct topics while low-diversity passages centred on one discourse topic. With frequency held constant and the target words counterbalanced across conditions, it was possible to measure how contextual history influenced word learning. In Experiment 1, words experienced in the low-diversity condition outperformed words in the high-diversity condition, both in terms of learning word forms and their meanings. This pattern was also seen in Simulation 1, with words appearing in low-diversity networks receiving more activations. Informed by the results of Simulation 2, Experiment 2 was designed to first anchor words onto one familiar topic before introducing variation. Relative to frequency-matched words that continued to be experienced in the same discourse topic, words seen in a new topic were better learned in terms of form (but no clear difference in the learning of meaning). Findings from the two experiments show that it is possible to manipulate context and induce variation in word learning in a single laboratory session. Clearly, such learning is some distance from the complex item-level knowledge about a word that builds slowly over time via natural language experience. Nevertheless, the advantage of the current approach is that it allows the learning environment to be carefully controlled and the effects of one variable to be measured.

Low semantic diversity and anchoring

In contrast to Johns et al. (2016), Experiment 1 found that novel words experienced in low-diversity contexts were advantaged both in terms of learning word form and their meanings. A similar effect has also been seen in word learning experiments with toddlers (e.g., Horst, Parsons, & Bryan, 2011; Williams & Horst, 2014). One explanation, following Horst's (2013) work with children, is that more similar contexts help channel attention from comprehending the texts to processing the novel words, thereby boosting word learning. It is not clear that this explanation best fits the low-

diversity advantage seen in Experiment 1. While the discourse topic remained the same in the low-diversity condition, the passages actually varied from episode to episode across both conditions (i.e., not repeated verbatim); note also performance on the comprehension questions was high and equivalent across conditions, and the discourse topics were familiar to participants.

A different, yet related, type of explanation stems from the idea of anchoring. On this view, when experiencing novel words in texts of familiar topics, low semantic diversity may help secure the novels words onto pre-existing knowledge in long-term memory (Mak, 2019). This way, the nascent representations of the novel words may be stabilized and reinforced. A consequence of this may be greater prominence and/or accessibility in memory, in line with the learning advantage seen in Experiment 1.

The notion of anchoring can be conceptualized as clustering coefficient in network science. In Simulation 1, target words in the low-diversity networks (vs. those in the high-diversity networks) received more activation. Such an advantage was attributed to target words in the low-diversity networks having higher clustering coefficients, meaning that their immediate neighbourhood is denser. Higher density may facilitate activation to circulate back to the target nodes, thereby reinforcing the nascent representations of the target word (Goldstein & Vitevitch, 2014). The finding that the simulation result converged with behavioural data from Experiment 1 provides greater confidence that it was the contextual manipulation that drove the low-diversity in Experiment 1, not a third factor such as arousal.

Next, we consider the extent to which anchoring is psychologically plausible. Specifically, is there any evidence showing that high-diversity words are initially experienced in a restricted set of contexts? Given our experiments each took place in just one session, we are cautious not to make any bold claims with regard to how and when anchoring might happen in real life. However, it is possible to note some relevant observations. For instance, young children frequently ask their caregivers to repeat a storybook, or replay a video (Crawley, Anderson, Wilder, Williams, & Santomero, 1999; Sulzby, 1985; Yanek, 2018), and indeed, word learning from a story is enhanced when toddlers hear the same story verbatim (Horst et al., 2011; Williams & Horst, 2014; see also Childers, Paik, Flores, Lai, & Dolan, 2017; Scott & Fisher, 2012). Second, and more generally, language experience takes place in more restricted contexts early in development with diversity growing as children engage with more speakers and in different communicative contexts; learning to read is also likely to bring variation and diversity to a word's contextual history (e.g., Montag, Jones, & Smith, 2015; Nation, 2017). Future experimental and computational work is needed to capture lexical and contextual history and to evaluate the extent to which developmental trends overlap with our notion of anchoring.

Although initial learning may benefit from anchoring/low diversity, this finding doesn't fit with corpus-derived work, which consistently found a high-diversity advantage in visual word recognition (e.g., Jones et al., 2017). This suggests that the benefit of anchoring/low diversity may be restricted to the initial stages of word learning. Specifically, we proposed that the benefit of anchoring on subsequent processing may saturate (Jorge-Botana et al., 2017; Murray & Forster, 2004), meaning that repeated anchoring beyond a certain threshold may result in little or even no detectable effect on behavioural measures. This is particularly evident in the accuracy data from speeded recognition across Experiments 1 and 2. In Experiment 1, novel words in the low-diversity condition were anchored onto one topic six times, while in Experiment 2, they were anchored onto one topic eight times. Despite the extra anchoring (and hence extra exposure), mean accuracy in speeded recognition was almost identical (94.4% in Exp 1 vs. 94.1% in Exp 2).[3] This, therefore, provides credence for the notion that the effect of anchoring on word learning may saturate. Saturation may also apply to diversity, but its trajectory may be different. Future studies should use computational methods to compare the saturation trajectory for anchoring and diversity (see for example Jorge-Botana et al., 2017; Jones & Rowland, 2015).

Anchoring, coupled with contextual variation, can further boost the learning of word form

Simulation 2 and Experiment 2 presented high-diversity novel words in one discourse topic for five exposures before contextual variation was introduced; this produced better learning of form than the single-context condition. We suggest the anchoring opportunity helped stabilize the novel word representations. We must stress that given results from Johns et al. (2016), anchoring should *not* be interpreted as a pre-condition for the high-diversity advantage seen in Experiment 2; instead, it should be conceived of as a facilitator or a contributor, which enabled the novel words to be better placed to take advantage of contextual variation subsequently. One consequence of contextual variation may be greater context-independence (i.e., being less tightly bound with a specific context; Jones et al., 2012). In network science, greater context-independence can be conceptualized as being linked to more unique word nodes, which could further help activation to circulate back to the novel target when a relatively dense immediate neighbourhood is in place.

As others have argued, a behavioural marker of greater context-independence might be enhanced word identification in isolation (Hsiao et al., 2019) and in new contexts (Pagán et al., 2020). Relatedly, the Semantic Distinctiveness Model (Jones et al., 2012; Jones et al., 2017; see also Nelson & Shiffrin, 2013) states that when a word is experienced, the encounter will be compared to its contextual history stored in memory; if the current context is distinct from the word's contextual

[3] In the high-diversity conditions, mean accuracy increased from 89.3% (Exp 1) to 95.4% (Exp 2).

history, there is more pressure for stronger encoding (Bolger et al., 2008; Johns et al., 2016). What follows is that the learning opportunity afforded by an encounter in a diverse context is more valuable than one in a repeating or highly similar context (e.g., Jones et al., 2012; Kachergis, Shiffrin, & Chen, 2009; Lev-Ari, 2016; Nation, 2017). A related explanation is that the high-diversity words in Experiment 2 might have become more polysemous as they were experienced in an additional topic. Polysemous words in general are recognized faster than non-polysemous words in lexical decision (Rodd et al., 2002; 2004), plausibly due to the former receiving more supportive activation from their distinct but overlapping senses (see also Floyd & Goldberg, 2020). This explanation fits with the view that semantic diversity is a continuous measure of word senses (Hoffman & Woollams, 2015). However, Cevoli et al. (2020) recently challenged this view by showing that semantic diversity values derived from corpus analysis do not capture variation in meaning, but variation in topics. Future work is therefore needed to delineate what it means for a word to be polysemous and how this relates to contextual experience.

A retrieval-type account may also explain the high-diversity advantage seen in Experiment 2. It is well-accepted that retrieval practice benefits learning more than restudy (e.g., Hogan & Kintsch, 1971; McDaniel & Mason, 1985; Roediger & Karpicke, 2006; Tulving, 1967). In addition to greater encoding, the high-diversity condition might have provided additional retrieval opportunity, further assisting consolidation and learning (see also van den Broek, Takashima, Segers, & Verhoeven, 2018). In the anchoring passages, a novel word in the high-diversity condition was initially and strongly associated with one topic before being re-encountered in a new topic. Experiencing the same novel word in a different topic potentially introduces subtle shades of meaning and nuance. Participants may have been sensitive to these semantic subtleties such that they recalled the first topic while processing the second. This retrieval opportunity would serve to boost learning, relative to the low-diversity words whose meanings were within the same discourse topic throughout the experiment.

Finally, and turning to the learning of meaning, previous studies (Hoffman & Woollams, 2015; Hsiao et al., 2019; Johns et al., 2016; Jones et al., 2012) found that high-diversity words are more difficult to judge in relatedness tasks. In Experiment 2, however, the meanings of high-diversity novel words were judged equally well as their low-diversity counterparts. This could be because the high-diversity words were only experienced in two topics, suggesting that their semantic representations may not have grown complex enough to impact on performance in a semantic judgement task. Of note is that while relatedness judgement is a widely adopted paradigm in psycholinguistics, interpreting the results can be tricky in the context of word learning. If for example high-diversity novel words were judged more poorly (as in Experiment 1), two interpretations are

possible: (i) the meanings of the novel words were poorly learned (hence poor task performance), and (ii) the meanings of the novel words were learned fairly well but their complex contextual history made semantic judgement difficult. As such, future studies interested in how word learning is influenced by contextual variation should consider using a range of tasks to tap semantic knowledge to complement relatedness judgement (e.g., meaning generation; Bolger et al., 2008).

Pseudoword inferiority as an index of word learning

Alongside highlighting the utility of word learning experiments, our experiments introduced the pseudoword inferiority effect as an index of word learning. Building from the premise that words generate a top-down influence that can interfere with letter identification, if a novel word has become well-established, it should induce a pseudoword inferiority effect. Evidence for this was seen in both experiments, showing that six to eight exposures to a new form in incidental reading is sufficient to drive changes in performance on the pseudoword inferiority task. The size and direction of the effect was also sensitive to our contextual manipulation. In Experiment 1, novel words encountered in the low-diversity condition showed more interference than those in the high-diversity condition. In Experiment 2, the direction of the effect flipped, with high-diversity words bringing about more interference. These findings from the pseudoword inferiority task mirror those seen in speeded recognition, highlighting the utility of the pseudoword inferiority task in tapping word learning and recognition.

Word learning from context: some reflections

Word knowledge exists on a continuum—from very weak and impoverished (e.g., a newly learned word) to very accessible and complex (Horst, 2013) and each encounter with a word—be it in spoken discourse or in text—offers an opportunity to refine its representations (Nation, 2017). Not surprisingly, words that are encountered more frequently tend to be more accessible than words that rarely occur in natural language (e.g., Brysbaert, Mandera, Keuleers, 2018; Nelson & Shiffrin, 2013). Our work adds to the evidence base showing that in addition to frequency, the context of the encounter influences verbal learning and memory, and the semantic content matters in the conceptualization of contextual history (Jones et al., 2017). Experiment 1 demonstrated that diversity in context may not be beneficial to the initial stages of word learning. This may be because the weak representations that characterise newly encountered words benefit from stabilization via more similar (and topically familiar) contexts rather than enrichment via diverse contexts. In contrast, Experiment 2 showed that given an anchoring opportunity, further enrichment via diverse contexts supports word learning.

At present, evidence from learning experiments is somewhat mixed. Some studies found diverse contexts to be more beneficial for word learning (e.g., Frances, Martin, & Duñabeitia, 2020; Johns et al., 2016; Rosa et al., 2017); others reported mixed results (e.g., Wilkinson & Houston-Price, 2013), no effect (e.g., Joseph & Nation, 2018), or a negative effect (e.g., Horst et al., 2011; Williams & Horst, 2014). This inconsistency is not surprising as there are numerous factors that may mediate the effect of contextual variability, including, but not limited to, age of participants (toddlers vs. undergraduates), topic familiarity (familiar vs. unfamiliar topics), word class of the target words (noun vs. verb), training paradigm (intentional vs. incidental), duration of training, and whether the non-diverse condition was implemented as verbatim repetition (Bolger et al., 2008; Pagán & Nation, 2019). Data from longer term word learning experiments are also needed to bridge to understanding precisely what a corpus-derived metric such as semantic diversity reflects.

Limitations and conclusions

Our focus has been with investigating the consequences of manipulating contextual variability on word learning while other important factors such as frequency were held constant. Any difference between the high and low diversity conditions within each experiment cannot be a simple consequence of differences in frequency as this was equated across conditions. However, there was a frequency difference *between* experiments. In Experiment 1, each novel word was seen six times (once in six different passages) whereas in Experiment 2, each word was seen once in eight different passages. The amount of variation also differed across the two experiments. In Experiment 1, high-diversity contexts were sampled from six different discourse topics whereas in Experiment 2, variation came from switching the discourse topic once. This was an active decision to keep the experiment within a single session and therefore not complicated by multi-session learning. Ultimately, of course, word learning is typically an incremental and dynamic process spanning a prolonged period of time. Future studies should consider multi-day training paradigms and the effect of number of discourse topics, not only to tackle more words and more varieties of variation, but also to consider how these factors interact with other factors known to influence learning and memory (e.g., spacing and testing effects) as well as sleep-related consolidation (e.g., James, Gaskell, & Henderson, 2019; Henderson, Weighall, & Gaskell, 2013). Relatedly, we measured learning immediately after exposure; there is a clear need to investigate longer-term retention via delayed post-tests. Finally, while both Simulation 2 and Experiment 2 suggested that anchoring helped bring about the emergence of a high-diversity advantage in the learning of form, our experimental design does not allow us to conclude that anchoring must precede diversity. Further work is required to determine the importance of learning order, although note that starting small (i.e., restricting

diversity) facilitates learning in different domains including word learning (Jorge-Botana et al. 2017) and aspects of morphology (Tamminen, Davis, & Rastle, 2015) and syntax (Elman, 1993).

In conclusion, Experiments 1 and 2 manipulated the semantic diversity of novel words while controlling for frequency. The two behavioural studies clearly indicated that item-level contextual history encoded during independent reading influenced lexical processing at the post-tests. Experiment 1 and Simulation 1 suggested that low semantic diversity (and topic familiarity) allowed the novel words to be anchored onto pre-existing concepts in long-term memory, thereby facilitating initial word learning. Experiment 2 and Simulation 2, on other hand, showed that if stabilized via anchoring, novel words can benefit from semantically diverse contexts, which further boosted lexical access.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity not 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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

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*, 255–278.

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

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019348

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. https://doi.org/10.1080/01638530701792826

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bouchet-Valat, M. (2019). SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. R package version 0.6.0. https://CRAN.R-project.org/package=SnowballC

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

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. https://doi.org/10.3758/BRM.41.4.977

Cervetti, G. N., Wright, T. S., & Hwang, H. J. (2016). Conceptual coherence, comprehension, and vocabulary acquisition: A knowledge effect? *Reading and Writing*, *29*(4), 761–779. https://doi.org/10.1007/s11145-016-9628-x

Cevoli, B., Watkins, C., & Rastle, K. (2020). What is semantic diversity and why does it facilitate visual word recognition? *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01440-1

Chan, K. Y., & Vitevitch,M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*,1934–1949. https://doi.org/10.1037/a0016902

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive Science*, *34*, 685–697. https://doi. org/10.1111/j.1551-6709.2010.01100.x

Childers, J. B., Paik, J. H., Flores, M., Lai, G., & Dolan, M. (2017). Does Variability Across Events Affect Verb Learning in English, Mandarin, and Korean? *Cognitive Science*, *41*, 808–830. https://doi.org/10.1111/cogs.12398

Chiesi, H. I., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 275–290.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*,407–428. https:// doi.org/10.1037/0033-295X.82.6.407

Crawley, A. M., Anderson, D. R., Wilder, A., Williams, M., & Santomero, A. (1999). Effects of repeated exposures to a single episode of the television pro- gram Blue's Clues on the viewing behaviors and comprehension of preschool children. *Journal of Educational Psychology*, *91*(4), 630–637. https://doi.org/10.1037/0022-0663.91.4.630

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. http://igraph.org

Diakidoy, I. A. N. (1998). The role of reading comprehension in word meaning acquisition during reading. *European Journal of Psychology of Education*, *13*(2), 131–154. https://doi.org/10.1007/BF03173086

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Floyd, S., & Goldberg, A. E. (2020). Children Make Use of Relationships Across Meanings in Word Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *2*(999). https://doi.org/10.1037/xlm0000821

Frances, C., Martin, C. D., & Duñabeitia, J. A. (2020). The effects of contextual diversity on incidental vocabulary learning in the native and a foreign language. *Scientific Reports*, *10*(1), 13967. https://doi.org/10.1038/s41598-020-70922-1

Freebody, P., & Anderson, R. (1983). Effects on text comprehension of different proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, *15*, 19-29. https://doi.org/10.1080/10862968309547487

Gaskell, M. G., Cairney, S. A., & Rodd, J. M. (2019). Contextual priming of word meanings is stabilized over sleep. *Cognition*, *182*, 109–126. doi: 10.1016/j.cognition.2018.09.007

Goldstein, R., & Vitevitch, M. S. (2014). The influence of clustering coefficient on word-learning: How groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, *5*(NOV), 2009–2014. https://doi.org/10.3389/fpsyg.2014.01307

Grainger, J., & Jacobs, A. M. (2005). Pseudoword context effects on letter perception: The role of word misperception. *European Journal of Cognitive Psychology*, *17*(3), 289–318. https://doi.org/10.1080/9541440440000131

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General*, *148*(2), 325–341. https://doi.org/10.1037/xge0000498

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*,662–720. https://doi.org/10.1037/0033-295X.111.3.662

Havas, V., Taylor, J.S.H, Vaquero, L., de Diego-Balaguer, R., Rodríguez-Fornells, A., & Davis, M. H. (2018). Semantic and phonological schema influence spoken word learning and overnight consolidation. *Quarterly Journal of Experimental Psychology*, *71*(6), 1469–1481. https://doi.org/10.1080/17470218.2017.1329325

Henderson, L. M., Weighall, A., & Gaskell, G. (2013). Learning new vocabulary during childhood: Effects of semantic training on lexical consolidation and integration. *Journal of Experimental Child Psychology*, *116*(3), 572–592. https://doi.org/10.1016/j.jecp.2013.07.004

Hills, T. T., Maouene, J., Riordan, B., Smith, L. B., & , L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273. https://doi.org/10.1016/j.jml.2010.06.002

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and related- ness effects in semantic tasks: Are they due to semantic coding? Journal ofMemory and Language, 55, 247–273. http://dx.doi.org/10.1016/j.jml .2006.04.001

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 385–402. https://doi.org/10.1037/a0038995

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.

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, *114*(June), 104146. https://doi.org/10.1016/j.jml.2020.104146

Horst, J. S. (2013). Context and repetition in word learning. *Frontiers in Psychology*, *4*(APR), 1–11. https://doi.org/10.3389/fpsyg.2013.00149

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*(FEB), 1–11. https://doi.org/10.3389/fpsyg.2011.00017

Hsiao, Y., Bird, M., Norris, H., Pagán, A., & Nation, K. (2019). The influence of item-level contextual history on lexical and semantic judgments by children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000795

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language*, *103*, 114–126. https://doi.org/10.1016/j.jml.2018.08.005

Jaeger, T. F. (2011, June 25). More on random slopes and what it means if your effect is not longer significant after the inclusion of random slopes [Blog post]. Retrieved from https://hlplab.wordpress.com/2011/06/25/more-on-random-slopes/

James, E., Gaskell, M. G., & Henderson, L. M. (2019). Offline consolidation supersedes prior knowledge benefits in children's (but not adults') word learning. *Developmental Science*, 22. https://doi.org/10.1111/desc.12776

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*, 278–305. http://dx.doi.org/10.1016/0010-0285(81)90011-6

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. https://doi.org/10.3758/s13423-015-0980-7

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, *73*(6), 841-855. https://doi.org/10.1177/1747021819897560

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across aging and bilingualism. *Frontiers in Psychology*, *7*(MAY), 1–11. https://doi.org/10.3389/fpsyg.2016.00703

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 67, pp. 239–283). Amsterdam, the Netherlands: Elsevier. http://dx.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), 115–124. https://doi.org/10.1037/a0026727

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Jorge-Botana, G., Olmos, R., & Sanjosé, V. (2017). Predicting Word Maturity from Frequency and Semantic Diversity: A Computational Study. *Discourse Processes*, *54*(8), 682–694. https://doi.org/10.1080/0163853X.2016.1155876

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, *166*(August 2017), 190–211. https://doi.org/10.1016/j.jecp.2017.08.010

Kachergis, G., Shiffrin, R. M., Yu, C., & Shiffrin, R. M. (2009). Frequency and Contextual Diversity Effects in Cross-Situational Word Learning. *Proceedings of the Cognitive Science Society*, *31*(31), 2220–2225.

Kaefer, T., Neuman, S. B., & Pinkham, A. M. (2015). Pre-existing Background Knowledge Influences Socioeconomic Differences in Preschoolers' Word Learning and Comprehension. *Reading Psychology*, *36*(3), 203–231. https://doi.org/10.1080/02702711.2013.843064

Kan, I. P., Alexander, M. P., & Verfaellie, M. (2009). Contribution of Prior Semantic Knowledge to New Episodic Learning in Amnesia. *Journal of Cognitive Neuroscience*, *21*(5), 938–944. https://doi.org/10.1162/jocn.2009.21066

Lev-Ari, S. (2016). How the Size of Our Social Network Influences Our Semantic Skills. *Cognitive Science*, *40*(8), 2050–2064. https://doi.org/10.1111/cogs.12317

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (Ed.), *Proceedings of the thirtieth Stanford child language research forum*. Stanford, CA: Center for the Study of Language and Information, 167– 178.

Mak, M. H. C. (2019). Why and how the co-occurring familiar object matters in Fast Mapping (FM)? Insights from computational models. *Cognitive Neuroscience*, *10*(4), 229–231. https://doi.org/10.1080/17588928.2019.1593121

Mak, M. H. C. & Nation, K. (2019, September). Consolidation of novel words into semantic memory: the role of context. Poster presented at ESCoP 2019. http://escop2019.webs.ull.es/wp-content/uploads/2019/09/ESCoP2019-ABSTRACTS.pdf

Mak, M. H. C., Hsiao, Y., & Nation, K. (2020, July 31). Lexical connectivity effects in immediate serial recall of words. https://doi.org/10.31219/osf.io/5yp4r

Mak, M. H. C., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired associate learning. *Psychonomic Bulletin & Review*, *27*, 1059-1069. https://doi.org/10.3758/s13423-020-01773-0

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*, 375–407. https://doi.org/10. 1037/0033-295X.88.5.375

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

McDonald, S., & Ramscar, M. (2001). The Influence of Context on Judgements of Semantic Similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Merced.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

Millard, S. P. (2013). EnvStats: An R Package for Environmental Statistics. Springer, New York.

Murray, W.S., & Forster, K.I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, *111*, 721–756.

Nagy, W. E., Herman, P. A., & Anderson, P. A. (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

Nelson, A. B., & Shiffrin, R. M. (2013). The Co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394. https://doi.org/10.1037/a0032020

Pagán, A., & Nation, K. (2019). Learning Words Via Reading: Contextual Diversity, Spacing, and Retrieval Effects in Adults. *Cognitive Science*, *43*(1), 1–24. https://doi.org/10.1111/cogs.12705

Pagán, A., Bird, M., Hsiao, Y., & Nation, K. (2020). Both Semantic Diversity and Frequency Influence Children's Sentence Reading. *Scientific Studies of Reading*, *24*(4), 356-364. https://doi.org/10.1080/10888438.2019.1670664

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Plunkett, K. (1998). Language Acquisition and Connectionism. *Language and Cognitive Processes*, *13*(2-3), 97-104. http://doi.org/10.1080/016909698386483

Pulido, D. (2007). The effects of topic familiarity and passage sight vocabulary on L2 lexical inferencing and retention through reading. *Applied Linguistics*, *28*(1), 66–86. https://doi.org/10.1093/applin/aml049

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The Mismeasurement of Mind: Life-Span Changes in Paired-Associate-Learning Scores Reflect the "Cost" of Learning, Not Cognitive Decline. *Psychological Science*, *28*(8), 1171–1179. https://doi.org/10.1177/0956797617706393

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 274-280.

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691619885860

Rodd, J., Gaskell, M. G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd, J., Gaskell, M. G., & Marslen-Wilson, W. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. https://doi.org/10.1016/j.cogsci.2003.08.002

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS ONE*, *12*(6), 1–12. https://doi.org/10.1371/journal.pone.0179004

Scott, R. M., & Fisher, C. (2012). 2.5-Year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*(2), 163–180. https://doi.org/10.1016/j.cognition.2011.10.010

Siew, C. S. Q. (2019). spreadr: An R package to simulate spreading activation in a network. *Behavior Research Methods*, *51*(2), 910–929. https://doi.org/10.3758/s13428-018-1186-5

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(3), 394–410. https://doi.org/10.1037/xlm0000139

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Sulzby, E. (1985). Children's emergent reading of favorite storybooks: a developmental study. *Reading Research Quarterly*, *20*(4), 458–481. https://doi.org/10.1598/RRQ.20.4.4

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003

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.

van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual Richness and Word Learning: Context Enhances Comprehension but Retrieval Enhances Retention. *Language Learning*, *68*(2), 546–585. https://doi.org/10.1111/lang.12285

Wang, H., Savage, G., Gaskell, M. G., Paulin, T., Robidoux, S., & Castles, A. (2017). Bedding down new words: Sleep promotes the emergence of lexical competition in visual word recognition. *Psychonomic Bulletin & Review*, *24*(4), 1186–1193. https://doi.org/10.3758/s13423-016-1182-7

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*, 59-85.

Wilkinson, K. S., & Houston-Price, C. (2013). Once upon a time, there was a pulchritudinous princess . . .: The role of word definitions and multiple story contexts in children's learning of difficult vocabulary. *Applied Psycholinguistics*, *34*(3), 591–613. https://doi.org/10.1017/S0142716411000889

Williams, S. E., & Horst, J. S. (2014). Goodnight book: Sleep consolidation improves word learning via storybooks. *Frontiers in Psychology*, *5*, 184. https://doi.org/10.3389/fpsyg.2014.00184

Wu, J. (2019). chinese.misc: Miscellaneous Tools for Chinese Text Mining and More. R package version 0.2.1. https://CRAN.R-project.org/package=chinese.misc

Yanek, D. (2018, December 12). Why do toddlers want to read the same book over and over again? [Blog Post]. Retrieved from https://www.todaysparent.com/toddler/toddler-behaviour/why-toddlers-want-to-read-the-same-book/

Appendix A: List of base words in Experiments 1 and 2

Experiment 1	Experiment 2
avidity	avidity
castigate	castigate
constellation	constellation
diffident	diffident
latent	latent
mendacious	mendacious
predilection	predilection
proliferation	proliferation
shenanigans	shenanigans
vagarious	vagarious
	architecture
	imbroglio

Appendix B: List of Pseudowords

List 1	List 2
allitive	abbusion
appurate	attusive
arregory	chesume
dratula	chrabble
flematic	draimant
genisty	glansact
lainstrom	lonbust
meciancy	mauranty
ossanic	prasmodic
plurious	splange
spredule	bropence
chyrene	chratum

The pseudowords in the shaded boxes were only used in Experiment 2.

Appendix C: Semantic associates of the base words.

Base words	Associate 1	Associate 2
Avidity	Enthusiasm	Passion
Castigate	Criticise	Reprimand
Constellation	Variety	Group
Diffident	Shy	Timid
Latent	Conceal	Hidden
Mendacious	Dishonest	Fraudulent
Shenanigans	Mess	Problem
Predilection	Inclination	Preference
Proliferation	Increase	Expansion
Vagarious	Erratic	Unpredictable
Architecture	Construction	Structure
Imbroglio	Humiliation	Embarrassment

Items in the shaded boxes were only used in Experiment 2.

Appendix D: Parametric settings for the spreadr simulations in Simulations 1 and 2.

Simulations 1 and 2 used the following parametric settings.

1. Start run: 50 (Simulation 1); 40 (Simulation 2a); 60 (Simulation 2b)

Start run refers to how much activation a particular node, usually the word of interest, initially receives at Time 0. In Siew (2019), an initial activation level of 100 was used to simulate lexical retrieval, false memory, and semantic priming. The lexical items in these paradigms were highly familiar words (e.g., *sleep*), so their representations must be very accurate and robust. When activated, these known words should have a significantly higher activation levels than the newly learned words in Experiments 1 and 2, which must have relatively impoverished representations. Therefore, the initial activation level for the novel words was reduced to 50 in Simulation 1 (where each target word had a frequency of 6), to 40 in Simulation 2a (where each target word had a frequency of 5), and to 60 in Simulation 2b (where each target word had a frequency of 8).

2. Retention: 0.1 to 0.9

Retention refers to the percentage of activation that remains in a node at each time step. Following Vitevitch et al. (2011) and Siew (2019), nine values were used: 0.1 to 0.9, in increments of 0.1. This was to ensure that "the simulation results were consistent across various values of retained level" (Siew, 2019).

3. Decay: 0.1

Decay refers to the proportion of "lost" activation at each time step of simulation. This parameter was not manipulated in Siew's (2019) simulation and was set at 0. However, given there exists empirical work (e.g., McKoon & Ratcliff, 1992) suggesting that spreading activation is a limited resource that decays over time, the current simulations attempted take this decaying property into account by setting the decay level to an arbitrary value of 0.1.

4. Suppress: 0.05

Suppress refers to the threshold where activation is suppressed to 0. This parameter instantiates the notion that nodes with extremely low activation levels are non-active and non-influential in spreading activation process (Siew, 2019). In her simulations, Siew kept this at 0, but this may not be appropriate for the Simulations here. Because firstly, there are significantly more nodes in the current networks (>200) than Siew's (~20), and secondly, some nodes in the current networks,

especially those that only co-occurred with the target word once in the reading passages, are expected to have very little or no influence on the spreading activation process. Therefore, the "suppress" parameter was arbitrarily set at 0.05.

5. Time step: 10

This parameter refers to the number of time steps over which the spreading activation process occurs. Following Vitevitch et al. (2011) and Siew (2019), this was set at 10.