Incidental learning of new meanings

for familiar words

Rachael Catherine Hulme

2018

Experimental Psychology

University College London

A thesis submitted for the degree of Doctor of Philosophy

Declaration

I, Rachael Catherine Hulme, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

………………………………………..

Abstract

Adults often learn new meanings for words they already know, for example due to language evolving with changes in technology (e.g., the newer internet-related meaning of "troll"). Learning new word meanings generally takes place incidentally, such as when reading for comprehension. The experiments in this thesis explore some of the different factors that impact adults' acquisition and long-term retention of novel meanings for familiar words learned incidentally from reading stories.

Experiment 1 assessed the effect of number of exposures on incidental learning. The results showed reasonably good memory of new word meanings after only two exposures, and a linear, incremental increase in recall with more exposures. There was also no forgetting after one week, regardless of the number of exposures during training.

Experiment 2 compared incidental to intentional learning, showing that new meanings for familiar words are harder to learn under incidental conditions, but may be less susceptible to forgetting. Experiments 3-4 explored whether a testing effect may have contributed to the good long-term retention of new word meanings in the previous experiments, and whether the method of immediate test affects this. These experiments showed that memory tests (cued recall or meaning-to-word matching) considerably enhanced retention of new word meanings.

Experiments 5-6 explored whether sleep is important for active consolidation of new word meanings, as previously shown for learning new word forms. In these experiments sleep improved explicit knowledge of new meanings when it occurred in the immediate interval between learning and test. No evidence of active consolidation was found; the results are consistent with a passive benefit of sleep in protecting against interference.

Together these experiments demonstrate that adult readers are proficient at learning new meanings for familiar words from a small number of encounters within naturalistic story contexts, and certain factors can have an important impact on learning.

Impact statement

Vocabulary learning is of critical importance because it has long-term consequences for academic attainment and employment in later life. Reading is a key source of new vocabulary from late childhood onwards, and adults learn the majority of new words and their meanings incidentally from context. The experiments in this thesis investigate vocabulary learning within a naturalistic story reading setting, which is important to better understand how adults typically learn new word meanings in everyday life. The research in this thesis focuses in particular on the learning of new meanings for familiar words. Only a few previous studies have investigated this, however ambiguous words are known to present a particular problem to those with comprehension difficulties (such as those with autism spectrum disorder). The research presented in this thesis contributes towards a better understanding of the processes involved in learning new word meanings within a typical adult population, which provides a useful basis for future research into developing interventions for those who have difficulties with vocabulary learning.

The present research has some specific implications for future academic research on vocabulary acquisition. The studies in this thesis highlight the possibility of using naturalistic texts in vocabulary learning research, which is not the case for many psycholinguistics studies of vocabulary learning. It is important to readdress the balance so that more research investigates learning under naturalistic conditions, as incidental and intentional learning conditions can have differential effects on acquisition and long-term retention of word meanings. Furthermore, the experiments reported in Chapter 3 of this thesis provide an important reminder that tests are not only tools for assessment, but also provide important opportunities for additional learning. Studies on vocabulary learning that compare memory between multiple test sessions should consider the influence that tests have on learning when interpreting the effects of additional factors, such as overnight sleep.

This thesis also has some practical applications for how to optimise language learning for first or second language learners. The findings show that learning new vocabulary from context is most effective with a larger number of exposures. Furthermore, testing memory immediately after reading enhances long-term retention of newly-acquired vocabulary. Such tests could be easily administered either by a teacher or through self-assessment using various test methods, such as cued recall or multiple-choice, and tests can enhance vocabulary learning even without any feedback on performance. Finally, language students can benefit from a night's sleep best if they read over revision notes just before bedtime the night before an exam.

Statement of contributions

Two authors created the short stories that were used to present the stimuli to participants in the experiments in this thesis. *Pink Candy Dream* was written by Helen Moss (a published children's author and former psycholinguistics researcher), and the other three stories (*Prisons*, *Reflections upon a Tribe*, and *The Island and Elsewhere*) were written by Johan Heemskerk (an unpublished student author).

Two UCL undergraduate students contributed to the work presented in this thesis. Dasha Barsky assisted with preparing the stimuli, collecting the data, and coding participants' responses for Experiment 1 in Chapter 2. Rachel Jose devised some of the comprehension questions for the stories, and assisted with coding the data for Experiment 2 in Chapter 3.

A modified version of Chapter 2 was accepted for publication in Language Learning prior to the completion of this thesis.

Acknowledgements

Contents

List of figures

List of tables

Chapter 1: General introduction

1.1 Introduction

Word learning in the native language (L1) continues throughout the adult lifespan. As well as frequently learning entirely new words and their meanings, adults must often learn new meanings for words already present in their mental lexicon. This occurs, for example, due to language evolving, especially due to changes in technology, e.g., the newer internet-related meaning of "troll" as a person who posts deliberately antagonising comments online. Adults can also encounter new meanings for familiar words when learning about a new subject or activity (e.g., the sailing term "boom" for a part of a yacht; Rodd et al., 2012), or when joining a new social or geographical community (e.g., the Scots dialect word "piece" meaning a sandwich).

Semantically ambiguous words, such as the previous examples, are commonplace in language. As many as 80% of English words are ambiguous (i.e., have more than one definition; Rodd, Gaskell, & Marslen-Wilson, 2002), and previously unambiguous words often acquire new meanings. This makes language comprehension more complicated, as ambiguous words can require additional processing compared to unambiguous words in order to ensure that the appropriate meaning of the word is selected in any given situation. Importantly, research has shown that adults make use of linguistic information learned throughout their lifetimes when interpreting the meanings of ambiguous words (Rodd et al., 2016; Wiley, George, & Rayner, 2018). Adults' representations of the meanings of ambiguous words depend upon their long-term experiences with the words' usage, but these representations are also altered based on recent experiences (Rodd et al., 2016), and so learning a new meaning for a word will affect how that word is processed in the future.

In contrast to the learning of new word forms and their meanings which has been widely researched (for reviews see Davis & Gaskell, 2009; Gaskell & Ellis, 2009), relatively little research has focussed on learning new meanings for familiar word forms. Some of the existing research has investigated children's learning of new meanings for familiar words (e.g., Casenhiser, 2005; Dautriche, Fibla, Fievet, & Christophe, 2018; Storkel & Maekawa, 2005; Storkel, Maekawa, & Aschenbrenner, 2013), and to date there have only been a few studies investigating this with adults, who have a more developed lexicon (Fang & Perfetti, 2017; Fang, Perfetti, & Stafura, 2016; Maciejewski, Rodd, Mon-Williams, & Klepousniotou, 2018; Rodd et al., 2012). However, it is important to investigate adults' learning of new meanings for familiar words in order to understand the initial development of semantic ambiguity. That is, how new meanings are combined with existing representations to create new ambiguous words in the mental lexicon.

1.2 Previous research on learning new meanings for familiar words

Some of the existing research suggests that learning new meanings for already-known words may be easier than learning entirely new words, as attention is not divided between learning a novel word form and mapping a new meaning onto that word (Storkel & Maekawa, 2005; Storkel et al., 2013). However others have suggested that it may be harder to learn new meanings for familiar words due to competition between the old and new meanings (Fang et al., 2016; Maciejewski et al., 2018; Rodd et al., 2012). Furthermore, it has previously been shown that children are slower to learn these words (Casenhiser, 2005) as it is harder for them to learn one-to-many mappings between word forms and meanings than direct one-to-one mappings. It may also be harder to learn a new meaning for a word with an already well-established meaning than to learn the two meanings simultaneously, due to the need to inhibit the more active dominant representation for the pre-existing meaning of the word (Dautriche, Chemla, & Christophe, 2016). Fang and Perfetti (2017) argue that learning new meanings for known words is a two-phase process in which familiarity with the word form may facilitate initial learning with the first couple of exposures, while inhibition due to meaning competition comes into play later after subsequent exposures to the newly-ambiguous word. Overall, these studies highlight some of the added complexities involved in learning new meanings for familiar words as compared with learning entirely novel words.

Another factor that can affect learning meanings for familiar words is the relationship of the new meanings to the pre-existing meanings of the words. There are two types of semantic ambiguity that can arise in language: polysemy and homonymy. Polysemy is when words have multiple semantically related senses of the same underlying meaning (e.g., a computer "virus" is related in function to a medical "virus"; Rodd et al., 2012). Homonymy, on the other hand, is when words have multiple semantically unrelated meanings (e.g., the "bark" of a tree/dog) that arise by chance, and it is less common than polysemy (Rodd et al., 2002). Rodd et al. (2012) compared learning new semantically related meanings to learning new semantically unrelated meanings for words. They found that recall of the new meanings for the previously unambiguous words was better for the newly-learned polysemous meanings than for the homonyms, which were harder to learn. Participants also responded more quickly to the newly polysemous words than to the newly homonymous words in a lexical decision task (Rodd et al., 2012). These findings are consistent with those of previous studies showing that while polysemy facilitates word recognition, homonymy delays recognition due to

competition from semantically unrelated meanings (Rodd et al., 2002; Rodd, Gaskell, & Marslen-Wilson, 2004). This effect likely arises as words with multiple related senses have highly overlapping semantic representations that make it quicker to settle into the appropriate representation, while for words with multiple unrelated meanings, the mutually exclusive representations of both meanings are initially activated, with semantic competition between these meanings increasing the time needed for a single meaning to be settled on (Rodd et al., 2004). These same underlying mechanisms may explain why homonyms are harder to learn than polysemes (Rodd et al., 2012). Although rarer in language than polysemy, homonymy poses a unique and interesting challenge to the learner, as they must acquire a novel word meaning alone and map it onto a known word form, without support from the existing representations for that word. This thesis therefore focusses on the learning of homonyms, for which the new meaning is not semantically related to the already-known meaning of the word.

1.3 Incidental vocabulary learning

Learning new L1 word meanings in everyday life generally takes place incidentally by inferring the new meaning from the surrounding context (Batterink & Neville, 2011), rather than through intentional memorisation of definitions. Incidental vocabulary learning can be defined as the learning of words and their meanings unintentionally whilst engaged in another activity, such as reading for comprehension (Hulstijn, 2003); in contrast to intentional learning, which is the deliberate attempt to memorise words and their meanings.

Reading is an important source of new vocabulary for both children and adults. From mid-childhood onwards the majority of new words and their meanings are learned through reading. Indeed, reading experience and reading comprehension ability have been shown to predict vocabulary levels at various ages (Cain & Oakhill, 2011). The Matthew effect in literacy and vocabulary development describes the phenomenon whereby children with better reading comprehension skills are more able to acquire new vocabulary through reading, thus further improving their reading abilities, and therefore further widening the gap in performance between good and poor readers over time (Stanovich, 1986). Reading and vocabulary are therefore very closely interlinked from an early age.

Reading is uniquely beneficial for vocabulary acquisition because the diversity of vocabulary used in written contexts tends to be richer than in spoken language. The intrinsic enjoyment of reading fiction is likely also an important factor for increasing the amount of reading. In addition to the amount of reading, the diversity of reading material may be particularly important for building stable representations of the meanings of words across

semantically diverse contexts (K. Nation, 2017). Fiction in particular has been shown to be important for vocabulary development across the lifespan. Data from a large-scale online vocabulary test has suggested that reading fiction specifically is as important for native language vocabulary development as reading in general ("Reading habits," 2013). This is likely the case as fiction typically contains a wider range of vocabulary than non-fiction writing. It is therefore important to investigate incidental learning of new word meanings through reading, as this is the way in which most vocabulary is learned.

The experiments in this thesis use a story-reading learning paradigm which provides ideal training conditions with which to study incidental vocabulary acquisition from reading. The training method has good ecological validity: adults acquire new meanings for familiar words incidentally while reading or listening for comprehension, and fantasy and science fiction stories are often a source for novel concepts to be mapped onto existing words (e.g., a "grim" is a large black ghostly dog and omen of death in the *Harry Potter* series of novels by J. K. Rowling). The story-reading incidental learning procedure used in the present work adopts a combination of the naturalistic elements of studies that have used authentic texts as the stimulus material (Godfroid et al., 2017; Saragi, Nation, & Meister, 1978), and careful within-item experimental control over key aspects such as the number of exposures, similar to methods used by Batterink and Neville (2011) and Pellicer-Sánchez (2016). Four short stories were written by two authors specifically for use in the present research. The stories included novel, invented meanings for existing unambiguous English words (e.g., a "foam" is a type of "safe concealed within a piece of furniture"), with the novel meanings conveyed through the stories' narratives. The new meanings for the familiar words are therefore encountered incidentally within the stories that the participants read for comprehension with no instruction to memorise the new meanings of the words. This naturalistic approach has not previously been used to explore the incidental learning of homonyms, as previous studies looking at this have used more intentional and less naturalistic learning conditions (Fang & Perfetti, 2017; Fang et al., 2016; Rodd et al., 2012).

1.4 Outline of thesis

The aim of this thesis is to investigate adults' incidental learning of new meanings for familiar words. The story-reading paradigm described above will be used to provide naturalistic incidental learning conditions, similar to the circumstances under which adults might encounter new meanings for familiar words in everyday life. A total of six experiments and one pre-test were run, with data from 498 participants analysed in total. This thesis addresses

several key theoretical questions relating to the acquisition of new meanings for familiar words both during learning and following the initial encounter with a new meaning; the focus of the individual chapters is outlined below.

Chapter 2 explores the impact of number of exposures on incidental learning and long-term retention of new meanings for familiar words. The experiment in this chapter also validates the use of the new story reading paradigm as an appropriate method for studying vocabulary learning.

Chapter 3 addresses two key questions through three experiments. Experiment 2 examines how incidental learning of new meanings for familiar words from stories compares to acquisition under intentional learning conditions. Experiment 3 and Experiment 4 explore the circumstances under which immediate memory tests may affect long-term retention of newly-learned word meanings.

In Chapter 4, Experiment 5 and Experiment 6 investigate the potential role of sleep in learning new meanings for familiar words.

Finally, Chapter 5 provides a summary and discussion of the main findings from this thesis.

Chapter 2: Incidental learning from reading: The impact of exposures

2.1 Introduction

For incidental learning from reading, certain factors concerning how new words and their meanings are presented in the text can impact on subsequent learning and retention. One likely key factor is the number of exposures to new vocabulary items (P. Nation, 2015). The impact of the number of exposures on adults' incidental vocabulary learning from reading has mainly been investigated in the domain of second language (L2) learning (e.g., M. Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010; Rott, 1999; Waring & Takaki, 2003; Webb, 2007). There are relatively fewer studies looking at adults' incidental acquisition of new words and their meanings in L1. The aim of the present chapter is twofold. Firstly, to answer the theoretical question of how the number of exposures affects adults' incidental learning of new meanings for familiar words. Secondly, to provide initial data on participants' performance on the newly-developed incidental learning task, in order to guide the experiments in the rest of this thesis.

2.1.1 Incidental L1 vocabulary acquisition from reading

All the studies on adults' incidental L1 vocabulary learning from reading to date have been concerned with the learning of new word forms. This has either entailed participants learning foreign or non-word labels for already-known concepts (e.g., Batterink & Neville, 2011; Mestres-Missé, Càmara, Rodriguez-Fornells, Rotte, & Münte, 2008; Mestres-Missé, Rodriguez-Fornells, & Münte, 2007; Pellicer-Sánchez, 2016; Saragi, Nation, & Meister, 1978; Williams & Morris, 2010), or in a few cases learning new words along with their novel, foreign or artificial meanings (e.g., Godfroid et al., 2017; Henderson, Devine, Weighall, & Gaskell, 2015).

An early study on L1 vocabulary acquisition from reading was a highly naturalistic study that used an authentic text as the stimulus material (Saragi et al., 1978). In the study native English-speaking participants read the novel *A Clockwork Orange* by Anthony Burgess, which contains 241 words in the fictional slang register "Nadsat" that are repeated on average 15 times (range = 1-209). Participants were not aware they would be tested on their memory of the novel words, and were instead told that they would be given a comprehension and literary criticism test. When their memory of 90 novel words was tested several days later in a meaning-to-word matching test, there had been significant acquisition of the words (76% correct) just from reading the narrative (Saragi et al., 1978). The researchers also found a significant positive correlation between the number of times a word occurred in the novel and the number of participants who correctly recalled the meaning. Saragi et al. (1978) suggest

that the minimum number of repetitions required for words to be learned incidentally while reading is "somewhere around ten" (p.76). However, since this early study, research has revealed different factors that contribute to incidental vocabulary learning depending on differing properties of the words. Therefore, focussing on a specific threshold to ensure learning is less useful than characterising the impact of number of exposures under typical incidental learning conditions.

Studies with ecological validity remain highly valued in the study of incidental vocabulary acquisition (Spivey & Cardon, 2015). A new eye tracking study by Godfroid et al. (2017) investigated participants' incidental learning of 29 Dari words (an Afghani dialect of Farsi) and their meanings whilst reading part of the novel *A Thousand Splendid Suns* by Khaled Hosseini in English, which was either their L1 or L2. The number of exposures to the Dari words in the text ranged from one to 23. As well as monitoring eye movements during reading, subsequent vocabulary acquisition was assessed through surprise tests of word form recognition, meaning recall, and meaning recognition. There was modest vocabulary learning: participants reading in their L1 scored 31.4% correct on word form recognition, 32.7% on meaning recognition, and 12.2% on meaning recall (Godfroid et al., 2017). Importantly, number of exposures was the strongest predictor of successful acquisition, more so than the total reading time summed across exposures (Godfroid et al., 2017). The eye movement data revealed a non-linear decrease in reading times across exposures with significant cubic and quadratic effects.

Godfroid et al. (2017) and Saragi et al.'s (1978) studies demonstrated clear incidental learning in the highly naturalistic context of reading real novels. However they lack experimental control over the number of exposures to the target words, which varied greatly in these authentic novels. Crucially, in such highly naturalistic materials the number of exposures may well be correlated or confounded with other properties of the new word meanings, such as how central they are to the story's plot, and some items may be intrinsically easier or harder to learn than others. This therefore emphasises the need for experimental control of the number of exposures in a within-item design.

In contrast to the previously discussed research, several studies (Mestres-Missé et al., 2008, 2007; Williams & Morris, 2010) have examined the processing and acquisition of novel L1 words with only a few exposures, but in less naturalistic contexts such as short sentences. In their eye-tracking study, Williams and Morris (2010) measured acquisition of 12 non-words using a two-choice synonym recognition test after participants had read a single meaningful sentence for each item. Average performance on this simple task was only 62% (Williams & Morris, 2010). Using different online processing measures, Mestres-Missé and colleagues

carried out an ERP study (Mestres-Missé et al., 2007) and an fMRI study (Mestres-Missé et al., 2008) to investigate meaning acquisition from context across three exposures with Spanish participants reading in their L1. In the ERP study they found that after three exposures to 65 items in contiguous sentences, brain potentials to novel words were already indistinguishable from real words. Participants showed moderate learning on a word pair task: they correctly recognised 69% of new word meanings, and correctly rejected 67% of incorrect meanings (Mestres-Missé et al., 2007). The fMRI study (Mestres-Missé et al., 2008) revealed similar acquisition from three exposures to 50 items (69% correctly identified meanings; 44% correctly rejected meanings). These studies using online measures of reading therefore provide some evidence for inferring and acquiring meanings of novel words from just one or three exposures in sentence contexts. However, the strength of these learning effects and the extent to which they translate into acquisition success remains unclear as these studies used only very simple post-reading vocabulary measures, if at all.

A few studies have combined elements of the more ecologically valid studies with experimental control of the number of exposures to items by using customised stories written or modified specifically for this purpose (e.g., Batterink & Neville, 2011; Henderson et al., 2015; Pellicer-Sánchez, 2016). Batterink and Neville (2011) investigated native English speakers' semantic integration of new meanings for 26 non-words, which were derived from context during story reading across ten exposures. They modified stories to give exactly ten exposures to the target words and examined semantic integration using the N400 ERP component, a negative component occurring around 400ms after stimulus onset whose amplitude varies in inverse relation to a reader's expectation of the upcoming word in a sentence (Kutas & Federmeier, 2011). Batterink and Neville (2011) found a greater reduction in N400 amplitude, indicating more semantic integration, for non-words embedded in consistently meaningful contexts than for non-words occurring in inconsistent, meaningless contexts. This reduction was already visible from the second exposure to the words. Acquisition was assessed explicitly through recall and recognition tasks; accuracy in recognising the meanings of the novel words was 72.4%, and accuracy on cued recall of meanings was 63.8%.

Another recent study by Pellicer-Sánchez (2016) used a story that had been purpose-written for their study to present their stimuli to participants reading in L2 and an L1 control group. They monitored participants' eye movements as they encountered the meanings of six non-words, each appearing eight times throughout the narrative. They found that, for participants reading in their L1, when tested immediately after reading accuracy in recognising the correct spelling for the new words was 91.3%. Accuracy in recognising the meanings for those words in a multiple-choice word-to-meaning matching test was 86.6%, and accuracy in

cued recall of the meanings was 65.3%. The eye-tracking data showed that participants reading in their L1 read the novel words significantly faster after only the first encounter, and after eight exposures they were read similarly to real, known words (Pellicer-Sánchez, 2016). Longer overall reading times were also associated with higher performance on the vocabulary measures.

These studies have demonstrated incidental learning of new words and their meanings through reading a single text in L1, although with somewhat mixed success. However, vocabulary gains from the reading of a single text are likely different to incidental learning through more extensive reading. Several studies with L2 learners (M. Horst, 2005; Webb & Chang, 2015) have found larger vocabulary gains from reading multiple different texts than typically found through reading a single text. There are various reasons why the amount of vocabulary learning may be greater from reading multiple texts; for example, within a single text there are smaller intervals between individual exposures, whereas multiple texts give more spaced encounters that may be more beneficial for learning (Webb & Chang, 2015). Additionally, words read in multiple texts are likely encountered in more diverse contexts (K. Nation, 2017), which may enable readers to build more stable representations of the meanings of words. However, conversely, children have been shown to learn vocabulary better from being repeatedly read the same storybook, as compared with the same number of exposures across different storybook contexts (J. S. Horst, Parsons, & Bryan, 2011). Caution must therefore be taken not to overgeneralise from findings of incidental vocabulary learning from reading one individual text to reading in general.

The studies reviewed here varied in ecological validity from the most naturalistic that used authentic novels as the reading material without experimentally controlling the context of exposure (Godfroid et al., 2017; Saragi et al., 1978), to non-naturalistic studies in which participants read individual sentences with only a few exposures to novel words (Mestres-Missé et al., 2008, 2007; Williams & Morris, 2010). Some recent studies have attempted to find a balance between these approaches (Batterink & Neville, 2011; Pellicer-Sánchez, 2016). Several additional differences between these studies could account for variation in acquisition success (e.g., number of items to be learned, measures used to assess learning, and whether participants learn both a novel word form and meaning or a novel word to describe an already-known concept). Number of exposures was consistently found to be a strong predictor of acquisition success for new word forms and their meanings (Godfroid et al., 2017; Saragi et al., 1978). Of the different aspects of vocabulary knowledge (including receptive and productive knowledge of the word form, meaning, and usage; P. Nation, 2001), productive knowledge of word meanings (assessed through cued recall) was the most difficult to acquire (Batterink & Neville, 2011; Godfroid et al., 2017; Pellicer-Sánchez, 2016), and may therefore

require more exposures for successful learning. Little research has investigated the incidental learning of word meanings in isolation from the acquisition of novel word forms, as is the case in learning new meanings for familiar words. As previously mentioned in Chapter 1, the added complexities involved in acquiring a new meaning for a familiar but semantically unrelated word may mean that a greater number of exposures would be required for sufficient learning as compared with learning a novel word form and its associated meaning.

2.2 Experiment 1: Number of exposures

2.2.1 Introduction

Experiment 1 investigated the effect of the number of exposures on adults' incidental learning and long-term retention of new meanings for familiar words in L1. In the present study, participants encountered new word meanings through reading a single text: one of four short stories that had been specifically written for the present research, with the novel, invented meanings for existing unambiguous English words conveyed through the stories' narratives. The number of exposures was manipulated within-subjects and within-item: each story contained four words with novel meanings, which were each presented two, four, six, or eight times throughout the text, counterbalanced across participants. Participants' knowledge of the new meanings was assessed through cued recall of the new meanings when presented with the words, and cued recall of the word forms when presented with definitions of the new meanings. Participants' memory was tested both immediately (following a short filler task) and one week after training. It was predicted that participants' accuracy in recalling the novel meanings and identifying which of the meanings paired with each word would be very low for only two exposures, but would increase gradually with an increasing number of exposures to the words with their novel meanings. It was further predicted that there would be significant forgetting of the novel meanings after the one-week delay, but that there would be better long-term retention with a greater number of exposures.

2.2.2 Method

Participants

Sixty-four participants took part and were included in the study (age: $M = 31.9$ years, $SD = 9.2$, range = 18-47; 32 female). The participants were recruited through the website Prolific Academic (Damer & Bradley, 2014). All participants were monolingual native

speakers of British English who were paid £3 for their participation in the first session of the experiment and £1 for the second session one week later. Of the 64 participants who completed the first session, 52 completed the delayed test a week later (81.3%).

In addition to the 64 participants included in the study, 18 participants took part but were excluded from the study: 11 for not meeting the language background criteria, six for getting more than one multiple choice comprehension question wrong when reading the story, and one due to a technical issue. The excluded participants were replaced with new participants to obtain the total of 64 participants included in the study.

Materials

Novel word meanings

The stimuli consisted of 16 English nouns (see Appendix A for a list of the stimuli) with only a single meaning in the Wordsmyth dictionary (Parks, Ray, & Bland, 1998). While all of the words had only a single dictionary meaning, most had several different related senses of that meaning; that is they were polysemous but not homonymous. (See Table 1 for descriptive statistics of the stimuli in each of the stories).

Novel concrete noun meanings were chosen to be semantically unrelated to the original meanings of the words[1], which was confirmed by a pre-test (see below); previous research has found that semantically unrelated meanings are more difficult to learn than semantically related ones (Rodd et al., 2012). Thirteen of the novel meanings were adapted from the stimulus set used by Rodd et al. (2012), and three additional meanings were devised following the same specifications. The new meanings were designed to be semantically diverse and consisted of hypothetical innovations ($n = 5$), natural phenomena ($n = 2$), invented objects ($n = 2$), social phenomena/traditions ($n = 5$), a technical term ($n = 1$), and a colloquial term ($n = 1$). Each of the new meanings had three distinguishing characteristic features, in order to maintain a similar level of complexity for each new concept. One sentence was written for each of the stimulus words to give a definition of the new meaning (e.g., "A foam is a safe that is incorporated into a piece of furniture with a wooden panel concealing the key lock, and each is individually handcrafted so that no intruders are able to recognise the chief use of the furniture."; see Appendix A for the full list). Each definition sentence incorporated the three

---

[1] The new meanings were created by swapping around pairs of words from a larger stimulus set of semantically related meanings (32 items in total, 16 of which were used in the present study). None of the previous semantically related meanings for the words were used in any of the stories.

key semantic features for the novel meaning (e.g., for "foam": "a safe inside a piece of furniture", "has a hidden key lock", and "individually handcrafted to fool intruders"), and the sentences were matched for length ($M = 32.9$ words, $SD = 3.7$). These sentences were given to the authors of the stories to be incorporated into story narratives. Abbreviated versions of these definition sentences were also written for use in the test task in which participants were asked to recall the word forms that paired with the definitions (see Appendix D for a list of these shorter, paraphrased definitions).

Relatedness pre-test

To ensure that the new word meanings were semantically unrelated to the words' existing meanings, a pre-test was carried out using a separate group of 20 monolingual native British English-speakers (age: $M = 30.1$ years, $SD = 10.0$, range = 18-52; 11 females). They rated the relatedness of the novel meanings presented in the definition sentences to the real, existing meanings of the words that they knew. The stimuli for the pre-test were the sentences giving definitions of the new meanings, each paired with a semantically unrelated word form. Each of the new meanings was also paired with a semantically related word form from a larger set of items not used in the present research[2] (e.g., "slot" for "a safe that is incorporated into a piece of furniture with a wooden panel concealing the key lock, and each is individually handcrafted so that no intruders are able to recognise the chief use of the furniture."). While none of these semantically related word-meaning pairs were used in the present research, these provided a frame of reference on the 7-point scale (where 1 indicated "highly unrelated" and 7 indicated "highly related"). The pre-test was split into two versions, with participants pseudo-randomly and evenly assigned to one of the two versions so that they saw each new meaning only once, paired with either the semantically unrelated or related word form. There were therefore ten data points for each meaning rated with its intended unrelated word. The results showed that, as intended, the 16 word form-meaning pairs used in the present study were perceived as unrelated to the existing meanings of the words (rating: $M = 1.8$, $SD = 0.3$, range = 1.3-2.6).

---

[2] All 32 meanings from the larger set of stimuli were included in the relatedness pre-test: the 16 items used in the present research, and 16 additional items not included in the present research. Rating data are given only for items included in the present study.

	n	Frequency (per mil.)	Frequency (log-transf.)	Orthographic Neighbourhood	Number of Letters	WordNet Senses	WordSmyth Senses	Age of Acquisition	Number of Semantic Associates	Semantic Relatedness Rating
Story 1	4	14.73 (12.26)	3.33 (0.44)	1.48 (0.21)	4.00 (0.00)	3.50 (2.08)	3.75 (1.89)	6.76 (0.83)	12.50 (5.80)	2.00 (0.61)
Story 2	4	14.70 (5.18)	3.45 (0.14)	1.59 (0.10)	4.25 (0.96)	5.75 (1.26)	6.50 (1.73)	7.27 (0.79)	16.50 (6.56)	1.63 (0.25)
Story 3	4	4.32 (5.18)	2.74 (0.46)	1.65 (0.41)	5.25 (1.50)	3.00 (1.83)	3.75 (3.10)	6.35 (0.68)	16.75 (7.68)	1.75 (0.13)
Story 4	4	21.30 (29.03)	3.28 (0.66)	1.04 (0.08)	3.50 (0.58)	4.00 (2.94)	3.75 (2.06)	6.05 (2.52)	14.50 (4.20)	1.75 (0.25)
All Words	16	13.76 (15.77)	3.20 (0.50)	1.44 (0.33)	4.25 (1.06)	4.06 (2.17)	4.44 (2.37)	6.61 (1.36)	15.06 (5.81)	1.78 (0.35)

Table 1. Descriptive statistics for the sets of stimuli in each of the stories. The means for each measure are displayed in the table, with standard deviations given in parentheses. The *n* refers to the number of stimulus words. The words frequency data reported are the SUBTLEX-UK word frequencies in occurrences per million and log-transformations of the raw word frequencies (log10[raw frequency+1]) (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The measure for orthographic neighbourhood is the OLD20 (orthographic Levenshtein distance 20) (Yarkoni, Balota, & Yap, 2008). Word sense data were taken from the WordNet (Fellbaum, 1998) and Wordsmyth (Parks et al., 1998) dictionaries. Age of acquisition data were taken from Kuperman, Stadthagen-Gonzalez, & Brysbaert (2012). The number of semantic associates counts come from Nelson, McEvoy, & Schreiber (2004). The semantic relatedness ratings refer to the results of the pilot study in which participants rated the relatedness of the stimulus words to their novel word meanings.

Short stories

Four separate stories were written, each incorporating four of the stimulus words in the context of their new meanings (see Appendix B for the stories). One of the stories (Story 1: *Pink Candy Dream*) was written by a professional children's author and former psycholinguistics researcher; the other three stories (Story 2: *Prisons*; Story 3: *Reflections upon a Tribe*; and Story 4: *The Island and Elsewhere*) were written by an unpublished student author. The authors were provided with a list of words with their novel meanings (the 16 items included in the present study, and 16 items not selected by them for inclusion in the stories), grouped broadly into four themes – one for each of the stories. They were asked to choose four of the items in each theme to incorporate into a story (selecting the items they felt would best fit together into a plausible narrative), with each word to appear eight times, providing information about its new meaning through the context. The stories were similar in length (Story 1: 2307 words; Story 2: 2320 words; Story 3: 2446 words; Story 4: 2330 words), and were designed to be similar in writing style and engaging for an adult audience. Each of the stimulus words appeared a total of eight times at naturally distributed positions within one of the four stories, with no stimulus word occurring in more than four consecutive sentences. The number of different words with novel meanings in each of the stories as a percentage of the total number of words was 0.2%. This is similar to the estimated percentage of novel "nadsat" words in *A Clockwork Orange* (0.4%; Saragi et al., 1978), indicating that the new word meanings were naturally-distributed and potentially learnable from the stories. On the first presentation of a stimulus word, sufficient information was given to allow the reader to derive the new meaning from the context right from the first exposure (e.g., "'Yes,' I murmured, breathing again. 'I *knew* it! It's a foam.' The ornate *chaise longue* was no ordinary piece of furniture, but concealed a built-in safe with an intricate key-operated locking system."). The amount of information about each new meaning in subsequent exposures varied naturally with the story narratives. None of the stimulus words appeared in any of the stories in the context of its real, existing meaning.

The short stories were then modified to vary the number of exposures to each stimulus word along with its novel meaning. Each of the four original stories contained eight exposures to each of the four stimulus words along with its novel meaning. The number of exposures was manipulated by removing some of the occurrences of the stimulus words in order to leave only two, four, six, or eight occurrences. This was achieved by replacing some of the instances of the stimulus word with words or phrases synonymous to the novel meaning (e.g., "foam" was replaced with "safe" or "hidden safe"), or in a few cases by simply omitting the word where it was not possible to use a synonym in the context of the narrative. This approach ensured that the amount of semantic content provided for each word was held constant

regardless of the number of exposures. In all of the exposure conditions the first and final occurrences of the stimulus word were kept in the story to minimise any primacy or recency effects; in the two-exposures condition these were the only occurrences. In the four and six exposures conditions, the additional occurrences of the stimulus words that were kept in were those appropriate to the natural narrative of the stories. In the eight-exposures condition all of the exposures were kept in. Each of the four stories contained one stimulus item in each of the four exposure conditions: two, four, six, and eight exposures, so that each participant saw an item in each of the conditions. Additionally, four versions of each of the stories were created so that each stimulus item appeared in each exposure condition across participants.

Design

Each participant read just one of the four stories. The independent variable of number of exposures to a word with its novel meaning was manipulated within-subjects and within-items: each participant was trained on four words that appeared two, four, six, and eight times respectively in the story. To ensure that each stimulus item was seen an even number of times in each exposure condition across participants, sixteen versions of the experiment were created (four per story). Participants were pseudo-randomly and evenly assigned to one of the sixteen versions of the experiment, with four participants assigned to each version. The independent variable of time of test (immediate versus one week later) was also within-subjects (based on the 52 participants who completed both sessions). The dependent variables measured were accuracy in cued recall of the novel meanings, and cued recall of the word form paired with each novel meaning.

Procedure

The experiment was conducted online using Qualtrics (Qualtrics, 2015), and was described to participants as "a study of different reading styles and the ability to understand texts". Participants were informed that they would be reading a short story and answering comprehension questions about what they had read, followed by a short vocabulary test and then some questions about their personal reading style. They were not made aware that they would encounter novel word meanings in the story, nor were they told to try to learn them, or that their memory for these novel word meanings would be tested. After completing the first session of the experiment, participants were not informed that they would be invited to complete a delayed test a week later. This was to discourage the use of deliberate

memorisation techniques by the participants, and to discourage rehearsal of the items over the week-long delay.

Each participant was pseudo-randomly assigned one of the four stories to read. Each story was divided into five pages of roughly even length and displayed on-screen one page at a time. After each page, a multiple-choice comprehension question appeared on a separate screen asking about details of the story's plot from the preceding page (without probing details of the novel word meanings). Participants were instructed to read the story closely and answer a question about what they had just read after each page; they were not given opportunities to re-read previous pages. Participants had to select the correct answer from four options (one correct), which appeared in a randomised order. The questions were designed to be very easy for any participant who had fully understood the text, participants were excluded if they got more than one of the five comprehension questions wrong, and as previously stated six participants were excluded on this basis.

After they had finished reading the story, participants completed a 34-item version of the Mill Hill vocabulary test (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996; Raven, Raven, & Court, 1998) as a filler task between the training phase and the testing phase. For each test item, participants were required to select one word from a list of six options that most closely matched the meaning of the presented word. None of the stimulus words appeared in the vocabulary test. The purpose of this task was to counteract any recency effects of memory for stimulus items encountered towards the end of the story.

Participants were then given a cued recall test of the novel word meanings that they had encountered in the story. Participants were presented one at a time with each of the four stimulus words they had encountered in the story and were asked to recall the appropriate novel meaning and type it into a blank text box. They were encouraged to provide as much detail as possible and to try to answer in full sentences even if they were unsure of their answer. If they could not remember anything about the new meaning for the word, they were instructed to type "don't know". For this test (and the subsequent test of cued recall of the word forms) the order of presentation of the items was randomised separately for each participant. Participants were only tested on the four items that had appeared in the story they read.

Participants were next given a cued recall test for the word forms that paired with each novel meaning. Participants were presented one at a time with short sentences that defined each of the novel word meanings. For each definition, participants were asked to recall the word that it described and type it into a blank text box. The definition sentences used for this test were abbreviated versions of the original definition sentences that were provided to the story authors. Although the sensitivity of this second test was expected to be reduced

compared to the initial test (due to priming of the word forms during the former test), it was included to provide a measure of memory that could be used in the event that participants were at floor on the initial test.

After completing both cued recall tests, participants provided demographics details, rated how enjoyable and clear they found the story on a 7-point scale, and answered questions about their reading style and habits. The primary purpose of these questions was to maintain the cover story that the purpose of the study was to investigate reading styles and comprehension, hence responses to these questions were not analysed.

Exactly seven days after the main experiment had been made available to participants, participants were invited to participate in a brief unexpected follow-up to the experiment. Participants began the delayed test an average of 7 days, 0 hours, and 45 minutes ($SD = 1$ hour 34 minutes, range = 6 days, 21 hours, 42 minutes–7 days, 5 hours, 15 minutes) after they had started the first session of the experiment. The delayed test session consisted of the same two cued recall tests, in the same order as in the first session, with the order of test items again randomised separately for each participant in both tasks.

2.2.3 Results

Analysis procedure

Responses for both cued recall tests were coded for accuracy by one of the experimenters (DB) blind to condition as either "1" for correctly recalled items or "0" for incorrect. The responses on the test of cued recall of the novel meanings were leniently coded as correct if at least one correct semantic feature was recalled (e.g., "a safe inside furniture" for "foam"). Any ambiguous or partially correct responses were resolved through discussion with another experimenter (RCH). The data were analysed with logistic mixed effects models, using the lme4 package (version 1.1-12; Bates, Mächler, Bolker, & Walker, 2015) and R statistical software (version 3.3.3; R Core Team, 2017). Four separate models were created: one for each of the two cued recall measures comparing accuracy between day one and day eight (which included only the participants who completed the tests at both time points, $N = 52$), and one for each of the two cued recall measures for all participants tested on day one only ($N = 64$). These latter analyses aimed to verify that the data from this larger set of participants did not differ from the subset who chose to complete both sessions.

The four models all contained random effects for participants and items (with slopes for exposure condition) and a fixed effect for exposure condition (four levels: two, four, six, or

eight exposures). The contrasts for this exposure condition factor were defined using orthogonal polynomial coding, with three separate contrasts to assess potential linear (two: -3, four: -1, six: 1, eight: 3), quadratic (two: 1, four: -1, six: -1, eight: 1), and cubic (two: -1, four: 3, six: -3, eight: 1) trends in the data. This approach was adopted as it is of greater theoretical interest to characterise the overall trend of the impact of number of exposures on acquisition of new meanings for familiar words, rather than using conventional contrasts to focus on differences between individual exposure conditions. The two models comparing performance between day one and day eight had an additional fixed effect for time, with the contrast defined using deviation coding (day one: -0.5 vs. day eight: 0.5), and a fixed effect for the interaction between time and the number of exposures (which was created by multiplying time by each of the contrasts for exposure condition). These models also included random slopes for time (i.e., day one vs. day eight) and the interaction between this variable and exposure condition by participants and items.

The first attempted model fit used the maximal random effects structure (as recommended by Barr, Levy, Scheepers, & Tily, 2013), which did not converge[3]. Following this, the models were simplified by removing only the correlations between the random slopes and random intercepts for the random effects by participants and items (without removing any of the random slopes). Three of the four models converged at this stage; the model comparing the data from day one and day eight for the cued recall of words measure did not converge. This model was simplified by instead removing the random intercepts by participants and by items, again leaving in all the random slopes (and this time leaving in the correlations between the random slopes), which allowed the model to converge. Therefore, all four analyses were carried out using models with simplifications of the maximal random effects structure as recommended by Barr et al. (2013).

Significance of the main effects and interactions was assessed using likelihood ratio tests by comparing the full model to identical models with only each factor or interaction of interest removed in turn (but leaving in any other interactions or main effects involving that factor or interaction), leaving the random effects structure intact. In the case of a significant effect of number of exposures, an additional analysis was run to determine whether there was a significant linear, cubic, or quadratic trend in the data. This was again assessed through likelihood ratio tests by comparing the full model to models with each of the components

---

[3] The "bobyqa" optimiser was used as per recommendations by Bates, Mächler, Bolker, and Walker (2016) for dealing with model convergence issues.

removed in turn. (The data and analysis scripts for this study are available at: https://osf.io/ybu6r.)

Cued recall of novel meanings

The data for accuracy in cued recall of the novel meanings comparing performance between day one and day eight ($N = 52$; see Figure 1) showed a reasonably high level of accuracy even after only two exposures (day one: 38.5%; day eight: 42.3%), appearing to increase in a positive linear trend with an increasing number of exposures. The data for the delayed test a week later showed the same pattern, and there appeared to be very little change in mean accuracy between these two time points. The analyses showed a significant main effect of number of exposures [$\chi^2(3) = 11.66$, $p = .009$][4], and no significant effect of time of test [$\chi^2(1) = 0.63$, $p = .429$], therefore showing no evidence of a difference in accuracy between the immediate test and the delayed test a week later. There was also no significant interaction between time and number of exposures [$\chi^2(3) = 1.58$, $p = .664$]. The trend analysis revealed that the number of exposures had a significant positive linear effect on cued recall of new meanings [$\chi^2(1) = 11.32$, $p < .001$], and no significant quadratic effect [$\chi^2(1) = 0.001$, $p = .973$] nor cubic effect [$\chi^2(1) = 0.15$, $p = .700$].

---

[4] Unfortunately it was not possible to obtain reliable measures of effect sizes (such as odds ratios and 95% confidence intervals) for the reported statistical contrasts as the LME model included a factor with more than two levels.
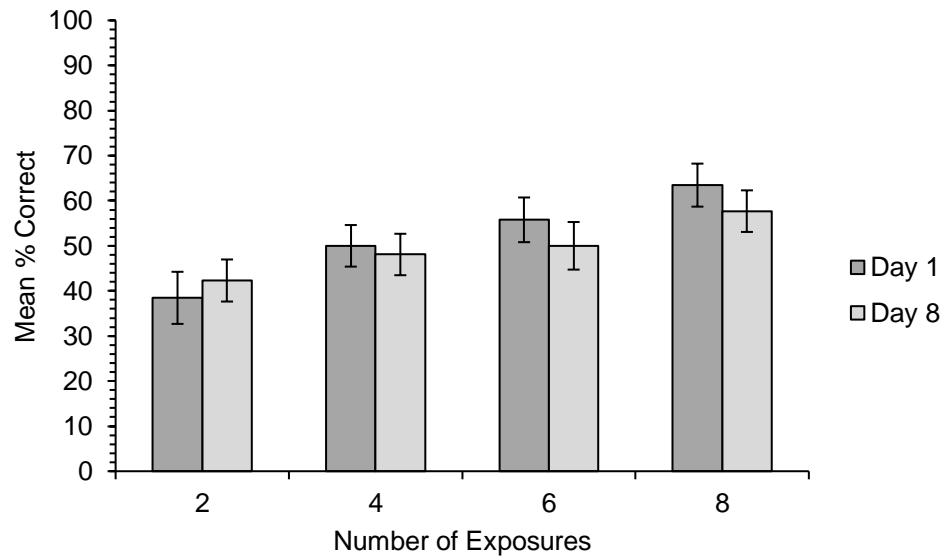
Figure 1. Experiment 1. Mean percentage of correct responses across participants for cued recall of novel meanings in each exposure condition when participants were tested on day one (immediately after training) and at the delayed test on day eight ($N = 52$)[5]. Error bars show standard error of the means, adjusted for the within-participant design (Cousineau, 2005).

The data for accuracy in cued recall of the novel meanings for all participants tested on day one ($N = 64$; see Figure 2) showed the same pattern as the data comparing performance between day one and day eight: a reasonably high degree of accuracy after only two exposures, which increased with an increasing number of exposures to the words with their new meanings. The results again showed a significant main effect of number of exposures [$\chi^2(3) = 11.12$, $p = .011$]. The trend analysis of the data also revealed a significant positive linear effect of number of exposures on cued recall of new meanings [$\chi^2(1) = 10.47$, $p = .001$], and no significant quadratic effect [$\chi^2(1) = 0.01$, $p = .929$] nor cubic effect [$\chi^2(1) = 0.65$, $p = .421$].

---

[5] The LME analyses were carried out on the raw binary accuracy data, however percentage data are displayed in the graphs for ease of interpretation.
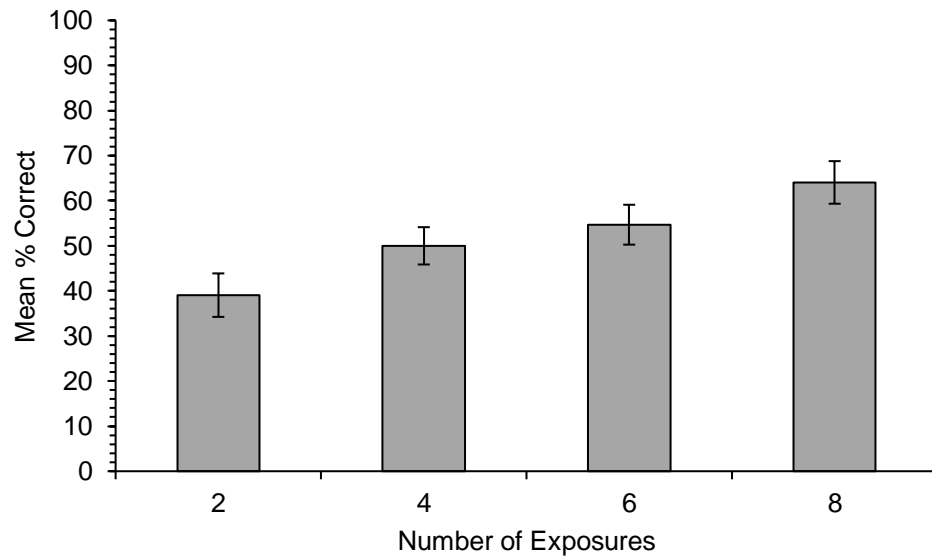
Figure 2. Experiment 1. Mean percentage of correct responses across participants for cued recall of novel meanings in each exposure condition for all participants tested on day one immediately after training ($N = 64$). Error bars show standard error of the means, adjusted for the within-participant design (Cousineau, 2005).

Cued recall of word forms

The accuracy data for cued recall of the word forms that paired with each of the novel meanings comparing day one to day eight ($N = 52$; see Figure 3) show that overall accuracy appeared to be higher in this test than in the cued recall of meanings test, although the pattern of the data appears broadly similar. These data again show a high level of accuracy after only two exposures (day one: 55.8%; day eight: 48.1%), with performance increasing gradually with a higher number of exposures. There was again very little change in accuracy between the tests on day one and day eight across all exposure conditions. The results showed that the main effect of number of exposures was marginal but non-significant for this measure [$\chi^2(3) = 6.82$, $p = .078$]. There was also no significant effect of time of test [$\chi^2(1) = 0.28$, $p = .599$], and no significant interaction between time and number of exposures [$\chi^2(3) = 0.99$, $p = .803$]. As the main effect of number of exposures was non-significant, any trends in the data were not assessed further.
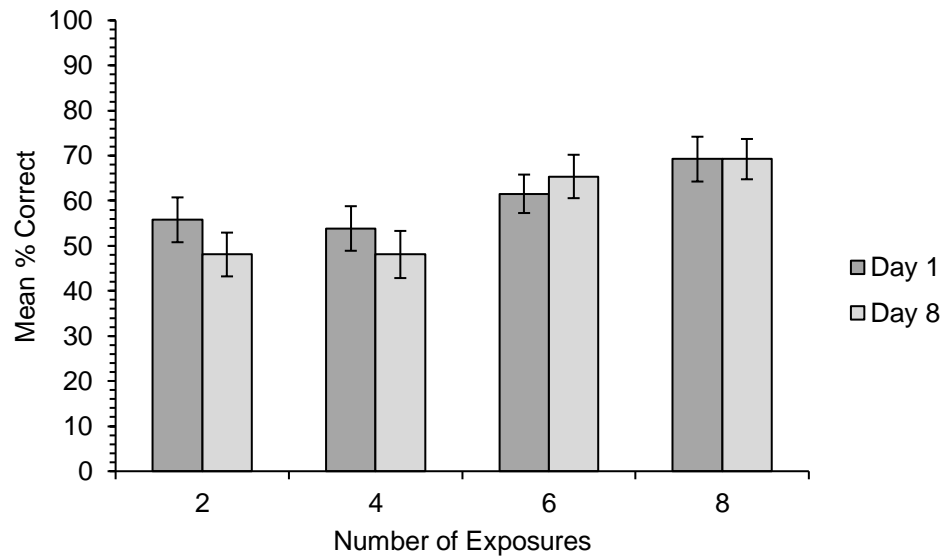
Figure 3. Experiment 1. Mean percentage of correct responses across participants for cued recall of word forms in each exposure condition when participants were tested on day one (immediately after training) and at the delayed test on day eight ($N = 52$)[6]. Error bars show standard error of the means, adjusted for the within-participant design (Cousineau, 2005).

The data for accuracy in cued recall of the word forms for all participants tested on day one ($N = 64$; see Figure 4) showed the same pattern. The results again showed no significant main effect of number of exposures [$\chi^2(3) = 3.95$, $p = .267$], so any trends in the data were not assessed further.

---

[6] The LME analyses were carried out on the raw binary accuracy data, however percentage data are displayed in the graphs for ease of interpretation.
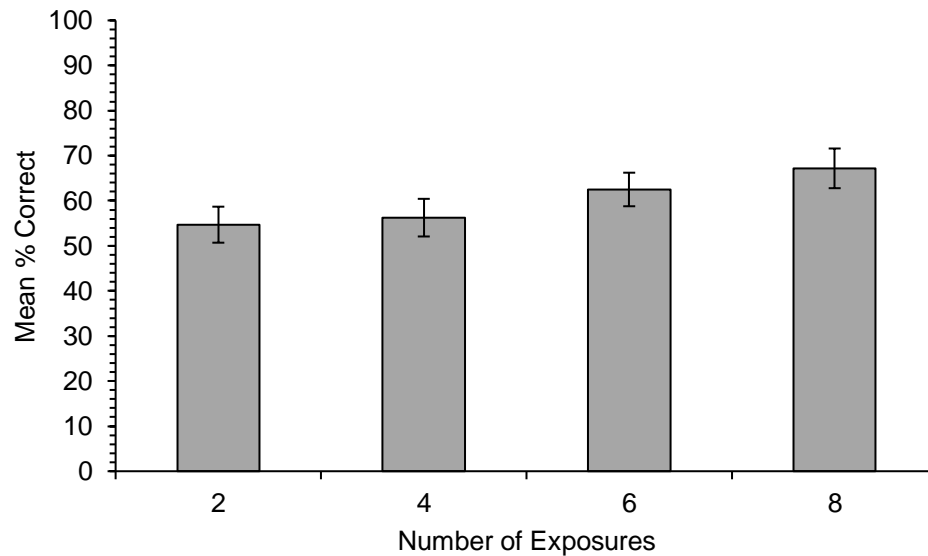
Figure 4. Experiment 1. Mean percentage of correct responses across participants for cued recall of word forms in each exposure condition for all participants tested on day one immediately after training ($N = 64$). Error bars show standard error of the means, adjusted for the within-participant design (Cousineau, 2005).

2.3 Discussion

The aim of Experiment 1 was to investigate whether adult readers can learn novel meanings for known words incidentally from stories after encountering very few instances of the novel word meaning, and how well these meanings are retained one week after exposure. Participants' memory of novel meanings for previously unambiguous words was assessed using tests of cued recall of the novel meanings and of the word forms that paired with definitions of the new meanings. The participants were tested both immediately after training and after a one-week delay.

Although there were substantial individual differences in performance, when tested immediately after training 38.5% of participants could correctly recall the new meaning for a known word after just two exposures in a single story context. These findings are consistent with some of the studies that used online measures to look at incidental learning of novel words and their meanings (Batterink & Neville, 2011; Mestres-Missé et al., 2007; Pellicer-Sánchez, 2016). Pellicer-Sánchez (2016) found that L1 participants read novel words that were embedded in a naturalistic story context significantly faster after only one exposure. The findings are also in line with the ERP studies of Batterink and Neville (2011) and Mestres-

Missé et al. (2007), which both showed evidence of semantic integration after only a couple of exposures to novel non-word labels for existing meanings.

Conversely, the present results are perhaps inconsistent with some of the behavioural measures of explicit memory for novel words and their meanings in previous studies. Both Williams and Morris (2010) and Mestres-Missé et al. (2008, 2007) found much higher accuracy in meaning recognition (66-69%) after only one or three exposures respectively. However, there are a number of differences between theirs and the present study that could account for the lower levels of acquisition we found. While in both Williams and Morris (2010) and Mestres-Missé et al.'s (2008, 2007) studies participants learned both the forms and meanings of a greater number of words than used in the present study, they did so from reading in the more constrained context of short sentences. In these previous studies participants had to acquire a new word form and map it on to a known concept which was easy to deduce from the sentences; this is quite different from the present study in which participants had to acquire a novel concept from a broader context and map it onto an already-known word form. Furthermore these previous studies used only very simple measures of meaning recognition, which Pellicer-Sánchez (2016) notes is much less difficult to acquire than productive knowledge of word meanings as measured through cued recall.

Perhaps the most comparable to the present study in terms of learning conditions and explicit measures of learning was that of Pellicer-Sánchez (2016). While Pellicer-Sánchez (2016) did not measure acquisition after different numbers of exposures, after eight exposures accuracy in cued recall of the meanings for novel words was 65.3% for participants reading in their L1. This is close to the level of meaning recall found in the present study with eight exposures (63.5%), suggesting that learning new meanings for familiar words may not be harder than learning new words and their meanings. However, participants in the (Pellicer-Sánchez, 2016) study were trained on more items (six) than in the present study (four), and with the same number of exposures to all items. Further research is therefore required to compare the acquisition of homonyms and non-homonyms directly within a single study.

Furthermore, as was predicted, the number of exposures influenced learning, with a linear increase in performance on cued recall of the new meanings with an increasing number of exposures to stimuli in the written text. The data for the cued recall of word forms measure showed roughly the same trend, although no significant main effect of number of exposures was found. (This was most likely due to performance on this second task having been enhanced by priming effects from the presentation of the word forms in the prior test of cued recall of the new meanings, although no feedback was provided to participants on either of the tasks.) The finding of a significant overall effect of number of exposures is consistent with

previous studies on incidental learning of word forms and their meanings, where number of exposures was shown to be a strong predictor of learning (Godfroid et al., 2017; Pellicer-Sánchez, 2016).

Importantly, in the present study the trend analyses for the significant effects of number of exposures on cued recall of the new meanings show that within the exposure range tested here, recall accuracy increased linearly as the number of exposures increased. As previously mentioned, recall accuracy at the immediate test was reasonably good, at 38.5% after only two exposures. However, the percentage increase in recall accuracy for each subsequent increase of two exposures was not nearly as high as that attained for the first two exposures. There was a steady incremental increase of 8.3% on average with each additional two exposures up to a maximum of 63.5% accuracy with eight exposures. The large difference between recall accuracy for the initial two exposures and the much smaller average increase for each subsequent two exposures suggests that the first one or two exposures are especially important for the acquisition of homonyms. The findings of previous eye-tracking studies (Godfroid et al., 2017; Pellicer-Sánchez, 2016) suggest that this may be because more time is spent reading and processing the initial exposures.

These results suggest that the initial couple of exposures have a disproportionally large impact on learning, while subsequent exposures all have a similar, lower level of impact. The positive linear pattern in the data likely arises due to a gradual dilution of the contribution of the initial exposures with an increasing total number of exposures. (Although see Bisson, van Heuven, Conklin, & Tunney, 2014, for an alternative explanation of similar findings.) However, had we tested larger numbers of exposures it is likely that learning gains would eventually plateau, similar to the pattern seen in the eye-tracking and ERP studies (Batterink & Neville, 2011; Mestres-Missé et al., 2007; Pellicer-Sánchez, 2016) where processing of novel words became indistinguishable from processing of known words after a few exposures. Within the relatively limited range of exposures tested in the present study though, acquisition of the new homonyms showed a steady linear increase with increased exposure. Based on previous research comparing the learning of homonyms to polysemes (Rodd et al., 2012), we would predict that the incidental learning of new semantically related meanings for known words would be even easier than learning new semantically unrelated meanings as in the present study. The initial exposures may have an even greater impact on the learning of polysemes due to support from the existing representations for the word's meaning; learning gains would also likely plateau after fewer exposures than for learning homonyms.

It is important to note that the learning gains seen in the present study are specific to the reading of a single text, as opposed to multiple texts. Some studies of L2 learning have found

higher levels of vocabulary acquisition from more extensive reading (M. Horst, 2005; Webb & Chang, 2015) than usually reported in studies of learning through a single text. This may be due to several contributing factors, such as increased spacing between encounters, and greater contextual diversity of individual exposures (K. Nation, 2017). The stimuli in the present study were highly contextually constrained within the stories; it is likely that incidental learning of homonyms would be more successful if encounters were distributed across separate stories. Further research is required to explore learning new meanings for familiar words through reading multiple texts, which would help build a clear picture of how adults typically learn L1 vocabulary.

Perhaps most surprisingly, in contrast to the predictions, participants showed no significant forgetting of the new meanings at a retest one week later (as shown on both measures), and long-term retention was not differentially affected by the number of exposures. None of the previously mentioned studies assessed long-term retention for participants reading text in their L1 (Batterink & Neville, 2011; Godfroid et al., 2017; Mestres-Missé et al., 2008, 2007; Pellicer-Sánchez, 2016; Saragi et al., 1978). However, Pellicer-Sánchez (2016) retested some of their group of proficient L2 learners in the same study following a two-week delay. They also found no significant forgetting between the immediate and delayed tests on measures of meaning recall, meaning recognition, and form recognition.

In contrast, another study in which intermediate L2 learners read a level-appropriate English novel, Waring and Takaki (2003) found that memory for novel words decreased in general after one week and had drastically decayed after three months. Contrary to the present study, they also found that words with a greater number of exposures were more resistant to forgetting over time. However, there are considerable differences in the learning conditions of these previous studies (Pellicer-Sánchez, 2016; Waring & Takaki, 2003) in which participants read and learned new words in their L2, as participants' general L2 vocabulary knowledge would have undoubtedly impacted on acquisition success. The vast differences between these studies and the present study in which participants read and learned new meanings in their L1 therefore make direct comparisons difficult.

A possible explanation for the maintained levels of recall accuracy seen over the course of one week concerns the testing effect (e.g., Roediger & Karpicke, 2006). This describes the phenomenon whereby the inclusion of a memory test immediately following training can facilitate long-term retention due to extra retrieval practice giving a boost to learning, even in the absence of any feedback on performance. In the present study the immediate tests could (even in the absence of feedback) have boosted performance on the delayed test. However, as Pellicer-Sánchez (2016) also notes, participants did not encounter the stimuli between the two

test sessions and they were not aware of the retest beforehand so had no cause to rehearse the stimuli during the preceding week. The results are therefore still a good indication of the long-term retention of new meanings for familiar words one week after incidental acquisition. The possibility that immediate memory tests may facilitate long-term retention of new meanings for familiar words will be explored in detail in Chapter 3.

A limitation of the present study is that sufficient information was included to elucidate the new meaning for a word on the first exposure. While this may happen sometimes in authentic texts, this is often not the case, and the amount of contextual information provided in individual exposures has been shown to influence vocabulary gains for L2 learners (Webb, 2008). However, this was necessary in the design of the present study to ensure that the key semantic information was available in all of the exposure conditions. Since the stories were custom-written by authors specifically for use in the current study, this allowed for complete experimental control over the number of exposures to the stimuli through the narrative in a within-items design. Importantly, this also allowed for control over potentially correlated or confounding factors such as the centrality of target items to the story's plot and properties of the words.

In conclusion, Experiment 1 extends what has previously been found in the L2 incidental vocabulary learning literature (e.g., Pellicer-Sánchez, 2016) to the learning of new meanings for previously unambiguous words in the native language. Some participants (38.5% at the immediate test) were able to successfully learn these meanings after just two exposures to familiar words with their novel meanings in a story context. Subsequent exposures additionally improved performance: learning increased linearly with an increase in the number of exposures in a cumulative incremental manner. Furthermore, knowledge of new meanings for known words was maintained well over the course of one week, regardless of the number of exposures during learning. Altogether, these findings demonstrate the remarkable success with which adults learn new meanings for known words incidentally whilst reading as in everyday life, as previously unambiguous words become homonyms.

Chapter 3: The testing effect in incidental learning of new meanings for familiar words

3.1 Introduction

Experiment 1 demonstrated the efficiency with which adults can acquire new meanings for familiar words incidentally through reading and retain them well over time. This naturalistic L1 vocabulary learning scenario in which adults acquire new word meanings incidentally during reading has seldom been compared against acquisition under intentional learning conditions that are more typically used in studies of adults' native language learning. The mode of learning may impact on efficiency of acquisition, but also long-term retention of the new word meanings. Another important factor that could influence people's ability to retain newly-learned word meanings over time is whether they are required to retrieve them during the intervening period between encoding and a later test. This chapter therefore explores two questions: how does incidental learning of new meanings for familiar words from stories compare to learning under intentional conditions, and what is the role of immediate memory tests in long-term retention of vocabulary learned in this way?

3.1.1 Incidental and intentional vocabulary learning

Incidental vocabulary learning is defined as learning words and their meanings whilst engaged in another activity such as listening or reading for comprehension (Hulstijn, 2003). In contrast, intentional vocabulary learning is defined as the deliberate attempt to memorise words and their meanings (Hulstijn, 2003). The incidental versus intentional learning dichotomy is often confounded with the overlapping concepts of implicit versus explicit learning in the literature on second language vocabulary acquisition, however these terms require distinction. Implicit and explicit learning are most clearly defined in terms of conscious awareness (DeKeyser, 2003): implicit learning is learning in the absence of online conscious awareness of what is being learned, while explicit learning is learning with conscious awareness. Intentional learning, involving deliberate memorisation of words and their meanings, necessarily requires awareness of what is being learned and thus makes use of explicit learning. However, it has been argued that incidental vocabulary acquisition, for example through reading, may involve both implicit and explicit learning processes (Rieder, 2003). It has further been posited (Ellis, 1994) that it is only word forms (and their grammatical usage) that can be acquired through implicit learning processes, while acquiring new word meanings and mapping word forms onto new meanings involve explicit learning mechanisms (Rieder, 2003).

The majority of studies on adults' vocabulary learning in their native language (e.g., Breitenstein et al., 2005, 2007; Fang & Perfetti, 2017; Fang, Perfetti, & Stafura, 2016; Perfetti, Wlotko, & Hart, 2005; Tamminen & Gaskell, 2013; Tamminen, Lambon Ralph, & Lewis,

2013; Van Der Ven, Takashima, Segers, & Verhoeven, 2015) have used training paradigms that encourage intentional learning. This is either done through directly instructing participants to learn words and their meanings, or less directly through the use of repetitive tasks that actively engage memory (e.g., associative learning paradigms). These types of word learning paradigms are favoured by researchers as they allow for training of a larger set of stimuli, due to the fast and efficient acquisition afforded by intentional learning procedures. However, this is not how the vast majority of native language words and their meanings are acquired (Batterink & Neville, 2011), and so conclusions drawn from such studies as to the nature of word learning and consolidation may not necessarily apply to how the majority of words and their meanings are acquired in everyday life (Henderson et al., 2015). Some of the studies on spoken word form learning have attempted to counteract intentional learning strategies through the use of incidental training regimes involving phoneme monitoring tasks (Dumay & Gaskell, 2007; Davis et al., 2009; Lindsay & Gaskell, 2013). However, such tasks are highly artificial as they only allow for learning of word forms and cannot be used in the study of word meaning acquisition.

A real-life context in which adults often learn new words and their meanings is when reading fiction, and this is especially true of science fiction stories. In light of this, several studies (outlined in Chapter 2) have adopted highly naturalistic methods in which adult L1 readers learn vocabulary from reading either authentic texts (Godfroid et al., 2017; Saragi et al., 1978) or texts modified or written specifically for the purposes of the studies (Batterink & Neville, 2011; Henderson et al., 2015; Pellicer-Sánchez, 2016), as is the case in the present research. In these studies participants read works of fiction with the primary focus being on comprehension, with vocabulary learning as a by-product. Additionally, to discourage intentional learning strategies, readers are not given any instruction to try to learn new vocabulary encountered in a text, and are not informed that their memory for the new words and their meanings will later be tested. The training methods used in these incidental learning studies are therefore much closer to the typical learning scenario for adults acquiring vocabulary in L1.

Most of the research comparing adults' incidental and intentional vocabulary acquisition comes from the literature on second language (L2) learning (e.g., Hulstijn, 1992; Lehmann, 2007; Peters, Hulstijn, Sercu, & Lutjeharms, 2009). However, Konopak et al. (1987) carried out a study in which eleventh graders (aged 16-17 years old) learned new subject-specific vocabulary from context in their native language. In the study an incidental learning group read a text passage about U.S. history which contained ten stimulus words and their meanings in context, while an intentional learning group read the same passage with the target words underlined and then completed a redefinition task in which they had to write a

definition for each of the words based on information in the text. Participants' vocabulary knowledge was assessed through a pre-test and a post-test (administered a day after training) in which they were asked to indicate if they knew the meaning for each given word or not, and to provide a definition of that meaning. They found that while there was some acquisition of knowledge by the incidental learning group, the intentional learning group showed greater vocabulary gains (Konopak et al., 1987). These findings are generally consistent with the consensus in the L2 vocabulary learning literature that intentional vocabulary learning is more efficient (e.g., Hulstijn, 1992; Peters et al., 2009), although others have found little benefit for intentional learning over incidental learning (Lehmann, 2007) or even that incidental learning from stories can be more efficient than combining a story with more focused study of L2 vocabulary (Mason & Krashen, 2004). However, it must be noted that success in incidental acquisition of L2 vocabulary through reading (or listening to) stories will depend greatly on participants' general level of proficiency in the language. Several recent studies with adult L1 readers have found good levels of native language vocabulary acquisition from reading alone (Batterink & Neville, 2011; Godfroid et al., 2017; Pellicer-Sánchez, 2016).

The conditions of initial vocabulary acquisition (incidental or intentional) prompt different types of information processing, which may also affect retention of words and their meanings. Vocabulary learned under intentional conditions may be retained better over time because of more attention being directly focussed on encoding the meaning for a word during training. Although incidental vocabulary learning from story reading may also have benefits for long-term retention, as new word meanings may be remembered better when presented in a rich and informative context (Webb, 2008). Additionally, it has been suggested that retention may be better for new word meanings inferred from context rather than those learned from given definitions, as the increased mental effort required at encoding is beneficial for later retrieval (Hulstijn, 1992). Indeed, Experiment 1 of this thesis showed that there was good retention of new meanings for familiar words learned incidentally through reading one week after training, regardless of the number of initial exposures.

Despite the fact that we learn the majority of new native language vocabulary words incidentally through reading from late childhood onwards, the majority of L1 word learning studies have focussed on explicit, intentional learning. Furthermore, little previous research has directly compared incidental L1 vocabulary learning through reading to vocabulary learning under intentional conditions. This is important because differences in processing novel word meanings under the different learning conditions at encoding may affect both how efficiently new meanings are acquired, and also how well they are retained over time. This chapter presents the first comparison between incidental and intentional learning of new meanings for familiar words.

3.1.2 The testing effect

Participants in Experiment 1 showed surprisingly good long-term retention of new meanings for familiar words after one week. In this experiment participants' memory of the new word meanings was tested immediately after training before being retested to assess long-term retention. This raises the question of whether the presence of the memory test immediately after training could have impacted on subsequent long-term retention. The testing effect describes the finding that testing memory following training can enhance long-term retention, as the additional retrieval practice at test affords an opportunity for further learning (for a review see: Roediger & Butler, 2011).

In vocabulary learning research, retrieval practice has been shown to lead to better retention of new words over time with adults learning L2 vocabulary (e.g., Fritz, Morris, Acton, Voelkel, & Etkind, 2007; Karpicke & Roediger, 2008; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013; van den Broek, Takashima, Segers, & Verhoeven, 2018) and children learning novel L1 words (Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014; Toppino & Cohen, 2009). For example, Karpicke and Roediger (2008) carried out a study in which English-speaking adult participants learned new foreign language (Swahili) words and their English translations in one of four conditions. The learning conditions consisted of either continuous studying and testing of all items, continuous studying of all items but with less testing, continuous testing of all items but with less studying, or lower amounts of both studying and testing. They found that accuracy in a cued recall test of the English translations for the words administered one week later was greatly increased for the conditions involving repeated testing, while repeated study showed no benefit (Karpicke & Roediger, 2008). This finding is consistent with many studies on the impact of retrieval practice on adults' L2 word learning (Fritz et al., 2007; Van den Broek et al., 2013, 2018), as well as some studies with children learning L1 vocabulary (Goossens, Camp, Verkoeijen, & Tabbers, 2014; Goossens, Camp, Verkoeijen, Tabbers, et al., 2014; Toppino & Cohen, 2009). The testing effect further enhances retention when feedback is provided on performance on the immediate test (e.g., Pashler, Cepeda, & Wixted, 2005), but retrieval practice is often beneficial even in the absence of any feedback (Roediger & Butler, 2011).

The testing effect is comparable to but distinct from the production effect and the generation effect that have been also described in the memory literature. The production effect is the phenomenon whereby words are remembered better when they are read aloud during encoding as opposed to being read silently (Ozubko & Macleod, 2010). The additional verbal

cue that is produced when reading the word aloud is helpful for later retrieval (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010; Sundqvist, Mäntylä, & Jönsson, 2017). The generation effect also concerns the benefit of production for memory, but at the retrieval stage rather than at the encoding stage. The generation effect describes the finding that information generated by participants (e.g., when identifying words in a word-fragment completion task, or identifying antonyms for word cues) is remembered better than information that is provided intact and only read by participants (Jacoby, 1978; Slamecka & Graf, 1978). The generation effect appears to be very similar to the testing effect, however the key distinction between the two phenomena lies in the mode of retrieval (Karpicke & Zaromb, 2010). In studies of the generation effect, participants retrieve information under incidental conditions without consciously thinking back to the past, on the other hand participants in studies of the testing effect are instructed to intentionally retrieve information from their episodic memory (Karpicke & Zaromb, 2010). Retrieval under intentional conditions, as is the case in the testing effect, produces a greater enhancement of future retention than generation of information under incidental retrieval instructions (Karpicke & Zaromb, 2010).

Despite the growing body of research on the benefits of retrieval practice for retention, the neurocognitive mechanisms underlying the testing effect remain somewhat unexplained (Antony, Ferreira, Norman, & Wimber, 2017). Influential models of word learning have failed to account for the testing effect. The Complementary Learning Systems (CLS) model of word learning describes how word forms are initially encoded into episodic memory in the hippocampus, and are integrated into semantic memory in the neocortex following a period of offline consolidation, such as during sleep (Davis & Gaskell, 2009; this is covered in detail in Chapter 4). However, this model does not account for the effect of conscious retrieval on memory for new words and their meanings. Antony et al. (2017) recently suggested that a similar mechanism may underlie both offline consolidation and the testing effect, which provides a fast track to consolidation. They argue that retrieval practice brings about the formation of flexible hippocampal-neocortical representations through the online reactivation of related knowledge (Antony et al., 2017). The testing effect is therefore important to consider in conjunction with offline consolidation processes to garner a full picture of how novel word meanings are remembered.

Retrieval practice has been shown to benefit long-term retention of information learned in different contexts, for example lists of vocabulary words (Karpicke & Roediger, 2008; Van den Broek et al., 2013), or learning from prose passages (Butler, 2010; Roediger & Karpicke, 2006a). However, little research has investigated whether the benefits of the testing effect differ for vocabulary acquired under different learning conditions. It is possible that retrieval practice may be more beneficial for novel vocabulary learned through incidental conditions

as learning under these conditions is more difficult, and testing after reading may focus participants' attention on the novel vocabulary items more directly. On the other hand, testing may be more beneficial for vocabulary learned under intentional conditions, as it could encourage more in-depth processing of new words and their meanings than is engaged under such training conditions. One study (Goossens, Camp, Verkoeijen, & Tabbers, 2014) has directly compared the impact of testing on children's learning of novel L1 vocabulary from a story context to learning new words in isolation. In the study, Dutch children aged 8-11 years learned difficult native-language vocabulary words either through listening to a story in which the new words appeared in context, or through listening to the words read out along with their synonyms in a list. Half of the items were then studied seven times, and the other half of the items were studied five times with two instances of retrieval (cued recall of meanings) interleaved. They were tested with cued recall of the meanings and a four-alternative forced choice (AFC) word-to-meaning matching test one week later. The results showed that children correctly recalled more meanings for words that had been tested rather than restudied during learning, although there was no difference in recognition. Additionally, children in the word list condition remembered the meanings for new words better overall than those in the story condition, and a marginally significant interaction showed that the testing effect was also slightly stronger for those in the word list condition. However, learning was not incidental in either condition in this study, and children who heard the story also had the meanings of the words explained to them in the study phases. It therefore remains to be seen whether the benefit of retrieval practice would differ for the learning of new word meanings acquired solely under incidental conditions in a story context, as compared with learning under intentional conditions.

One factor that may impact on the strength of the testing effect is the method of immediate test. The testing effect has been observed in studies using various different methods of immediate test, most usually with cued recall (for example with word pairs; Karpicke & Smith, 2012), but also with other test methods such as multiple-choice (e.g., Roediger & Marsh, 2005). There are several possibilities as to why certain methods of immediate test may be more beneficial for future retention than other test methods. The retrieval effort hypothesis states that testing is more helpful for long-term retention when it is more effortful (Pyc & Rawson, 2009). For example, in a study in which young adults learned the meanings of novel L1 vocabulary words, Karpicke and Roediger (2007) showed that increasing difficulty of retrieval by increasing the delay between initial study and initial testing (through cued recall) gave rise to better long-term retention than when initial retrieval effort was lower. Tests of productive vocabulary knowledge, such as cued recall of the word meaning, are more difficult than recognition tests in which new word meanings are supplied (Pellicer-Sánchez, 2016),

therefore an immediate test of cued recall of the word meaning may be more advantageous for future long-term retention that a multiple-choice recognition test.

On the other hand, immediate testing may be particularly beneficial when it assists with restructuring learned information into a format that is more helpful for long-term retention, which recognition tests may allow for. Multiple-choice recognition tests may aid retention due to the response choices cueing the retrieval of marginal knowledge that may otherwise not be easily accessible (Marsh, Roediger, Bjork, & Bjork, 2007). They may also provide an opportunity for additional learning of some items through the process of elimination of foils (Marsh et al., 2007) even in the absence of feedback on response choice. However, foil answers in multiple-choice tests may also lead to the learning of incorrect information (Butler, Marsh, Goode, & Roediger, 2006; Marsh et al., 2007; Roediger & Marsh, 2005).

Some studies have directly compared the effects of immediate cued recall and multiple-choice tests on long-term retention. For example, Duchastel (1981) investigated secondary school students' retention of a prose passage following testing with either a short-answer test (akin to cued recall), a multiple-choice test with three alternatives, or a free recall test without any cues. Students' memory of the prose passage was tested two weeks later with free recall of the topics discussed in the passage and a cued recall test asking a mixture of the same questions as on the immediate test, and questions either related or unrelated to these (Duchastel, 1981). Duchastel (1981) found that long-term retention (measured by cued recall) was better for those who had been immediately tested with the short-answer test, but no testing effect was observed for the other two groups. However, Duchastel (1981) found no testing effect for any group on the free recall measure of retention, and the cued recall measure of retention was a very similar format to the immediate test for the short-answer test group.

One concern about the testing effect is that information learned with the help of retrieval practice could be relatively inflexible and constrained, and may therefore not transfer to different types of delayed test. Tran, Rohrer, and Pashler (2015) and others have found that retrieval practice may not benefit later tests that require making deductive inferences about the learned information. Furthermore, Hogan and Kintsch (1971) found that immediate test methods that provide further exposure (i.e., recognition tests) are more beneficial than free recall for recognising items two days later, whereas tests involving retrieval (both free recall and recognition) boosted performance on delayed free recall. However, the degree to which different methods of immediate test aid future retention can differ depending on certain factors, such as the provision of feedback on performance. Kang, Mcdermott, Roediger, and Kang (2007) had participants study passages, followed by either a multiple-choice test, a short-answer test, reading statements, or a filler task (control). At a delayed test three days later,

they found that the group who had the immediate multiple-choice test performed better on tests of both multiple-choice and short-answer questions than the group with the immediate short-answer test (Experiment 1; Kang et al., 2007). However, in a second experiment where feedback was provided on initial test performance (Experiment 2; Kang et al., 2007), the group with the immediate short-answer test performed better on the delayed tests than those whose immediate test had been multiple-choice, supporting the retrieval effort hypothesis (Pyc & Rawson, 2009). Other studies have also found that the testing effect can transfer across different test methods (Butler, 2010; McDaniel, Anderson, Derbish, & Morrisette, 2007; Rohrer, Taylor, & Sholar, 2010), with cued recall quite consistently found to be more beneficial for long-term retention than recognition tests.

While deliberate retrieval practice has been found to enhance memory for diverse types of information, its underlying neurocognitive mechanisms remain unclear. Antony et al. (2017) have argued that the testing effect may involve a similar process as unconscious offline learning processes, providing a fast track to consolidation. The testing effect is therefore important from both a theoretical perspective to gain a comprehensive understanding of word learning, but also from a practical viewpoint as a way to enhance vocabulary learning. The testing effect has not previously been investigated in the context of the learning of new meanings for familiar words. The effect of retrieval practice is particularly interesting in this context, as testing participants' episodic memory of the novel meanings could allow them to construct new representations for the words more rapidly, which may partially compensate for the difficult learning conditions of acquiring an unrelated meaning for a familiar word. This chapter examines the impact of immediate retrieval practice on retention of new meanings for familiar words acquired under incidental and intentional learning conditions, as well as the efficacy of different test methods for enhancing their retention.

3.1.3 Chapter overview

The present chapter addresses two key questions on adults' learning of new meanings for familiar words through three experiments. The first question regards how learning new meanings for familiar words incidentally through the naturalistic story reading method (outlined in Chapter 2) compares with learning through a definition training task designed to encourage intentional learning strategies; this is addressed by Experiment 2. Memory for the new word meanings in this experiment is assessed through tests of cued recall of the new meanings for the words, and an eight-alternative multiple-choice meaning-to-word matching task both immediately after training and again one day (24 hours) later.

The second question addressed in this chapter concerns whether testing memory immediately following training improves subsequent long-term retention of new meanings for familiar words; Experiments 3 and 4 focus on two different aspects of the testing effect. Experiment 3 investigates the testing effect in learning under incidental and intentional learning conditions. Experiment 4 then examines the impact of the method of immediate test on the testing effect for learning incidentally through reading stories.

3.2 Experiment 2: Incidental versus intentional learning

3.2.1 Introduction

The question addressed by Experiment 2 was how the story-reading method designed for incidental learning of new meanings for familiar words would compare with a more intentional training task procedure. In Experiment 2 participants learned novel meanings for existing unambiguous words through both incidental story-reading (as in Experiment 1) and an intentional repetitive task-based learning procedure, with the same number of exposures to items learned through both methods (eight exposures). Their knowledge of the new meanings was then assessed first through cued recall, and second through a multiple-choice meaning-to-word matching recognition test, both immediately after learning, and again 24 hours later to assess long-term retention. There was no specific hypothesis as to which learning procedure would perform better, or whether the two training procedures would result in differential retention over time. However, it was hypothesised that both methods would show sufficient acquisition of the novel word meanings, as measured through accuracy in cued recall and recognition matching.

3.2.2 Method

Participants

Forty participants took part and were included in Experiment 2 (age: M = 30.1 years, SD = 7.1, range = 19-47; 23 female). All participants were monolingual native speakers of British English who were recruited through the website Prolific Academic (Damer & Bradley, 2014). Participants were paid for their participation in the first session of the experiment (£5) and additionally upon completion of the second session 24 hours later (£1). Of the final set of 40 participants who completed the first session, 31 also completed the 24-hour follow-up session

on time (77.5%), with one additional participant having been excluded from the second session due to completing it later than the pre-specified deadline.

In addition to the 40 participants included in the study, five additional participants took part but were excluded from the experiment: two due to not being monolingual native British English speakers, and three participants were excluded due to getting more than one of the multiple choice comprehension questions wrong in the story reading condition. The excluded participants were replaced with new participants to obtain the total of 40 participants included in the study.

Materials

Novel word meanings
The stimuli were the same 16 real English noun words with artificial new meanings as used in Experiment 1 (see Appendix A for the list of stimuli).

Short stories
The same four separate short stories as in Experiment 1 were used to present stimuli to participants in the incidental learning condition in this experiment (see Appendix B for the stories). The number of exposures to stimuli was not manipulated in the present experiment; each stimulus item appeared a total of eight times within a story.

Definition sentences
The sentences used in the initial definition reading phase of the intentional learning condition were the original sentences created to describe the three key semantic features of each of the novel word meanings (see Appendix A).

For the two-alternative multiple-choice task in the intentional learning condition, three shorter condensed, paraphrased versions of these sentences were created for each of the stimuli (length: $M = 11.29$ words; $SD = 2.13$). These abbreviated versions of the definitions omitted the word to which the meaning refers (e.g., for "foam": "A secure place to store valuables within an item of furniture."; "A safe with a wooden panel disguising the key lock."; and "A bespoke handcrafted piece of furniture containing a safe hidden from intruders."). Paraphrased versions were used in order to encourage the participants to read the whole sentence each time,

rather than relying on the simple recognition of the first word (see Appendix C for a list of the excerpts of the definitions used in this task).

One additional longer paraphrased version of each of the full definition sentences (used in Experiment 1 in the test of cued recall of word forms) was used in the eight-alternative multiple-choice test in this experiment (see Appendix D for a list of these paraphrased definitions).

Design

The experiment employed a within-subjects design: participants were trained on four items through the incidental learning condition and four items through the intentional learning condition. The independent variable of time of test (immediate versus 24 hours later) was also within-subjects (based on the 31 participants who completed both sessions). To ensure that each new word meaning was seen an even number of times in each condition, and that the order of the learning conditions was counterbalanced across participants, eight versions of the experiment were created. Half of the participants were trained on the items occurring in Stories 1 and 2, and the other half of the participants were trained on the items occurring in Stories 3 and 4; each participant was trained on half the total number of stimuli (eight items) as this was deemed to be a reasonable number of new meanings to learn in a single session. Within the set of items occurring in Stories 1 and 2, the words occurring in Story 1 were presented in the incidental learning condition (story) for half of the participants, with the words from Story 2 being trained through the intentional learning condition (definition training task), and vice versa for the other half of participants trained on the same set of items. The same organisation was used for the set of items occurring in Stories 3 and 4. Additionally, the order of the incidental learning condition and the intentional learning condition was counterbalanced across all participants in the experiment in an effort to minimise any order effects of the different tasks. Participants were randomly assigned to one of the eight versions of the experiment. The dependent measures were accuracy in cued recall of the new meanings for the words, and accuracy in the multiple-choice test.

Procedure

The experiment was carried out online using Qualtrics (Qualtrics, 2015), and participants were instructed to complete the full experiment in one sitting without breaks. Participants were not told to try to learn the word meanings in the experiment, and were not aware that their memory

for the new word meanings would be tested. After completing the experiment, participants were also not informed that they would be contacted at the same time the following day to invite them to complete the follow-up tests in order to discourage the use of deliberate memorisation strategies.

For the incidental learning condition participants read one of the four stories and answered the interleaved comprehension questions; the procedure for this task was identical to that of Experiment 1. Participants were excluded if they got more than one of the comprehension questions wrong: three participants were excluded on this basis.

The intentional learning condition consisted of two phases which both repeated once: definition sentence reading, followed by the two-alternative multiple-choice meaning-to-word matching task. In the definition reading phase, participants were presented with the original sentences that described the three key semantic features of each of the novel word meanings, stating the word to which it referred. These were presented one at a time on separate pages, and the order of presentation of the sentences was randomised for each participant. Participants were instructed to read each definition carefully to make sure they understood it before continuing on to the next one.

Once participants had read all of the definition sentences once, they moved immediately on to the two-alternative multiple-choice meaning-to-word matching task. In the task participants were presented one at a time with the three shortened, paraphrased versions of the definitions of each of the novel meanings. These shortened, paraphrased sentences omitted the words to which they referred, and for each item participants were instructed to choose the correct new meaning from two possible options: the correct word and one foil word. After selecting one of the options, participants were immediately provided with feedback on their choice on the same page, which either said "Correct answer!" or "Incorrect.". The items were presented in a pseudorandomised order, ensuring that the items referring to each word were roughly evenly-spaced, and none of the items referring to the same word occurred one after another. The foil word for each trial was one of the other words from the intentional training condition. Each foil word was paired an even number of times with each correct word, and the order that the correct word and foil word appeared in was randomised for each trial. The two phases of the intentional training were then repeated in the same order, with the order of presentation of items for the definition sentence reading randomised again for each participant, and the items in the two-alternative multiple-choice task presented in a different pseudorandomised order from the first time. This gave a total of two exposures to the novel word meanings from the definition sentence reading phase and six exposures to the new

meanings from the two-alternative multiple-choice task, giving a total of eight exposures, and equalling the number of exposures to the words in the incidental story-reading condition.

After they had completed training through both the incidental and intentional learning conditions, participants were given a brief filler task. This was the 34-item version of the Mill Hill vocabulary test (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996; Raven et al., 1998). The procedure was the same as in Experiment 1.

Participants were then asked to complete a cued recall test of all eight of the new meanings they had encountered in the experiment when presented with the words. The procedure was identical to the cued recall of novel meanings test used in Experiment 1. The order of presentation of the words was randomised for each participant, with the four words from each training method randomly intermixed within the test, and participants were only tested on the eight items that they had seen during training.

The second test task used in the present experiment was an eight-alternative multiple-choice meaning-to-word matching test. Participants were presented one at a time with the longer paraphrased versions of the definitions of the novel word meanings that they had been trained on. The sentences omitted the words to which they were referring, and for each novel meaning participants were asked to select the word that they thought matched the definition from a list of all eight of the stimulus words they had encountered throughout the experiment. The order of the eight words to choose from was randomised for each test item. The order of presentation of the new meanings was randomised for each participant. As with the cued recall test, participants were only tested on the eight items they had seen during training.

Finally, participants provided their demographics details and answered some questions about their reading habits. As with Experiment 1, these questions were used to maintain the impression that the experiment was investigating general reading and comprehension and the responses were not analysed.

Exactly 24 hours after the experiment had first been made available to participants, the participants were invited to take part in a short 24-hour follow-up to the experiment. Thirty-one of the original participants completed the follow-up tests, which they started an average of 24 hours and 1 minute (SD = 54 minutes, range = 22h26m-28h2m) after they had begun the first session the previous day. The follow-up tests consisted of a repeat of the two test tasks from the first session of the experiment: cued recall of the novel word meanings followed by the multiple-choice meaning-to-word matching task. The order of presentation of the items in each of the two tests was again randomised for each participant, and for the eight-alternative

multiple-choice test the order of the eight stimulus words to choose from was again randomised for each test item.

3.2.3 Results

Length of the training tasks

At the start of the different phases of the experiment (incidental learning condition, intentional learning condition, filler task, and test phase) participants were asked to report the time displayed on their computer in hours and minutes to give a rough guide as to the amount of time spent on each task. Participants spent significantly longer on average on the story reading task (including comprehension questions; $M = 12$ mins 30 secs, $SD = 4$ mins 34 secs) than the intentional training task ($M = 5$ mins 28 secs, $SD = 2$ mins 55 secs) $t(39) = 11.43$, $p < .001$.

Analysis procedure

Responses from the multiple-choice test were either coded as "1" for correct or "0" for incorrect with regards to which word had been selected to match with the meaning presented on each trial. Responses for the cued recall test were independently coded for accuracy by the experimenter (RCH) and a research assistant (RJ) blind to condition as either "1" for correctly recalled items or "0" for incorrect, as in Experiment 1. Any discrepancies between the two coders' decision on a given item were resolved on a case-by-case basis through discussion. One item ("bruise") was excluded from the analyses for the cued recall measure, as the percentage of participants who gave a correct response for that item in one of the two learning conditions (incidental, 20.0%) was less than two standard deviations below the grand mean for all items across both learning conditions ($M = 72.3\%$; $SD = 23.5$).

The binary accuracy data for the responses in both the multiple-choice and cued recall tests were analysed using logistic linear mixed effects models using the lme4 package (version 1.1-7; Bates et al., 2015) and R statistical software (version 3.0.2; R Core Team, 2017). Four separate models were used: one for each test measure comparing the results between day one and day two (including only the participants who completed both test sessions, $N = 31$), and one model for each measure for all participants tested on day one only ($N = 40$). These latter analyses aimed to verify that the data from this larger set of participants did not differ from the subset who chose to complete both sessions.

The contrasts for all of the factors were defined using deviation coding. All four LME models contained random effects by participants and items (with a slope for the random effect of learning condition by participants and by items[7]), and fixed effects for learning condition (incidental: -0.5 versus intentional: 0.5), position of the task in the experiment (first: -0.5 versus second: 0.5), and the interaction between learning condition and task position (which was created by multiplying the contrasts for these two factors). The two models for comparing performance between day one and day two contained additional fixed effects for time (day one: -0.5 versus day two: 0.5), and the interactions between time and learning condition, time and task position, and the three-way interaction. These models also included random slopes for time and the interaction between time and learning condition by participants and items.

To select the appropriate random effects structure for the model for each of the test measures, the first attempted fit used the maximal random effects structure (Barr et al., 2013). Only the model for all participants on day one for the meaning-to-word matching measure converged at this stage[8]. The other three models were then simplified by removing the correlations between random slopes and intercepts for the random effects by participants and items, without removing any of the random slopes themselves. The model for all participants tested on day one for the cued recall measure and both models comparing the results between day one and day two converged at this stage and were used as the final models for the analysis.

Significance of the main effects and interactions was assessed as for the previous experiment, through the use of likelihood ratio tests in which the full model was compared with identical models with only each factor or interaction of interest removed in turn (leaving in all other fixed effects and interactions, and keeping the same random effects structure).

To follow up on the significant interaction between learning condition and time for the multiple-choice meaning-to-word matching measure, simple effects analyses were done to analyse the effect of time separately for the two learning conditions. This was done by taking subsets of the data for the incidental and intentional learning conditions and creating a model for each containing fixed effects for time, training task position in the experiment, and the interaction, with random effects for time by participants and by items[9]. Significance of the

---

[7] Random slopes for the controlled factor of task position were not included in the models due to issues of model non-convergence. This factor was not of theoretical interest and the fixed effect for this factor was not significant in any of the models.

[8] The "bobyqa" optimiser was used as per recommendations by Bates et al. (2016) for dealing with model convergence issues.

[9] The models had the random correlations removed to match with the model used in the main analysis for this measure.

simple effects was determined using likelihood ratio tests to compare the full models to models without the fixed effect for time.

Cued recall of novel meanings
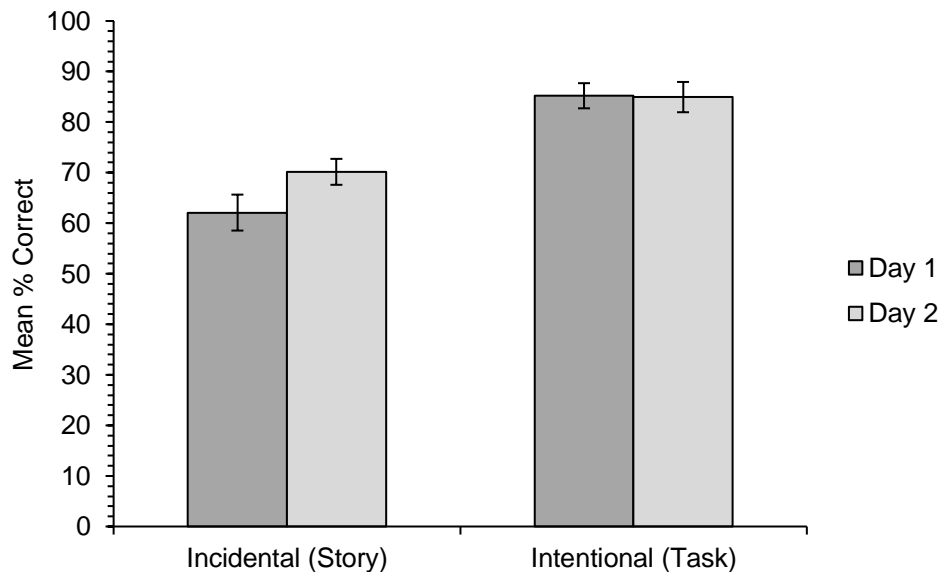
Figure 5. Experiment 2. Mean percentage of correct responses by subjects on the cued recall test[10] (meanings correctly recalled for the appropriate word) for each learning condition, when tested on day one (immediately after learning) and 24 hours later ($N = 31$). Error bars show standard errors for subject means, adjusted for the within-participant design (Cousineau, 2005).

The accuracy data for cued recall of the new meanings comparing recall between day one and day two ($N = 31$; see Figure 5) showed a reasonably high level of accuracy for items learned through both learning conditions, but with higher accuracy overall for items learned through the intentional learning condition (day one: 85.2%; day two: 84.9%) than those learned under incidental conditions (day one: 62.1%; day two: 70.1%). While accuracy improved slightly between day one and day two for items learned through the stories, accuracy remained at a similar level after 24 hours for items learned through the intentional condition. The analysis showed a significant main effect of learning condition [$\chi^2(1) = 14.32$, $p < .001$],

---

[10] The LME analyses were carried out on the raw binary accuracy data, however mean percentage accuracy data are displayed in the graphs for ease of interpretation.

and no significant main effect of time of test [$\chi^2(1) = 1.23$, $p = .268$], nor of position of the training task in the experiment [$\chi^2(1) = 0.83$, $p = .362$]. The interaction between learning condition and time was non-significant [$\chi^2(1) = 1.57$, $p = .210$], as were the interactions between learning condition and task position [$\chi^2(1) = 2.16$, $p = .141$], training task position and time of test [$\chi^2(1) = 1.79$, $p = .181$], and the three-way interaction [$\chi^2(1) = 0.83$, $p = .361$].
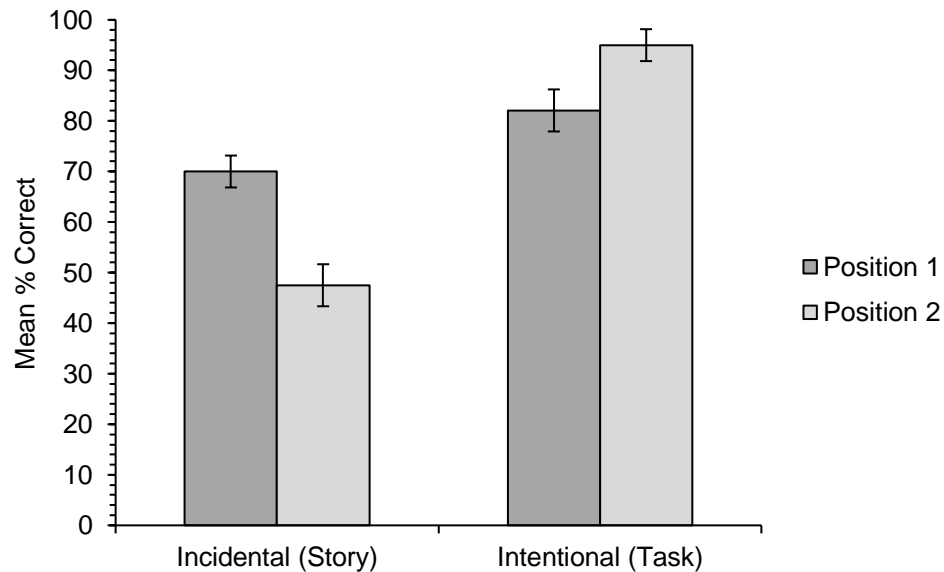
Figure 6. Experiment 2. Mean percentage of correct responses by subjects on the cued recall test (meanings correctly recalled for the appropriate word) for each learning condition, when that condition was presented in the first or second position in the experiment. The data presented are for all participants only for the session on day one immediately after training ($N = 40$). Error bars show standard errors for subject means, adjusted for the within-participant factor of learning condition (Cousineau, 2005).

The mean percentage accuracy data for all participants on day one only ($N = 40$) are shown in Figure 6. Again accuracy was significantly higher overall for items learned through the intentional training condition than the incidental condition [$\chi^2(1) = 21.35$, $p < .001$]. Again there was no significant main effect of task position [$\chi^2(1) = 0.0007$, $p = .979$]. Although there was a significant interaction between learning condition and task position [$\chi^2(1) = 4.68$, $p = .030$] whereby accuracy was higher for items learned through the stories when this had been the first task in the experiment (70.0%) than when this had been the second task (47.5%), whilst the opposite was the case for items learned through the intentional condition (first position: 82.1%; second position: 95.0%).

Multiple choice meaning-to-word matching
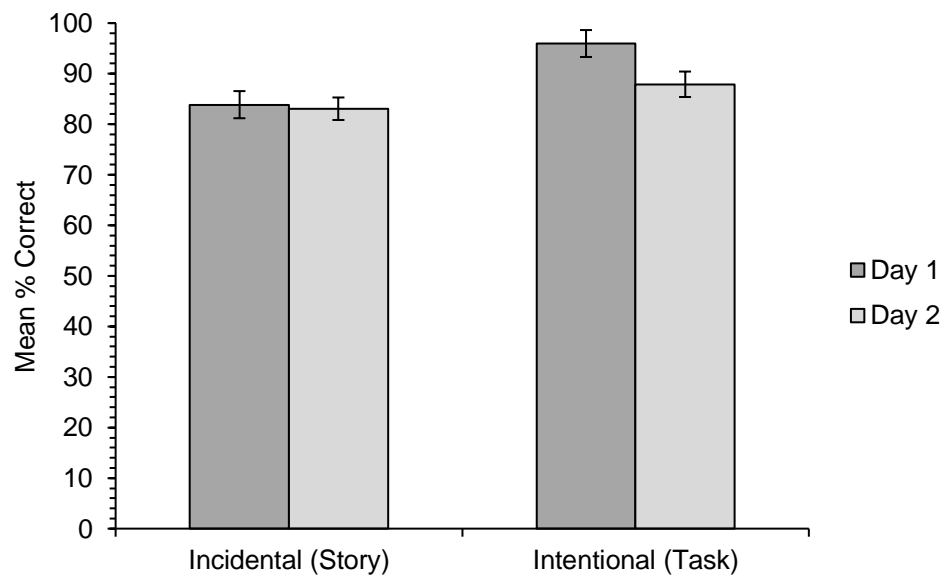
Figure 7. Experiment 2. Mean percentage of correct responses by subjects on the multiple-choice test (words correctly matched with the appropriate meaning) for each learning condition, when tested on day one (immediately after learning) and day two (24 hours later; $N = 31$). Error bars show standard errors for subject means, adjusted for the within-participant factor of learning condition (Cousineau, 2005).

The results for accuracy on the multiple-choice meaning-to-word matching test comparing results between day one and day two ($N = 31$) are shown in Figure 7. Accuracy was very high overall in both learning conditions, but was slightly higher for the intentional learning condition (day one: 96.0%; day two: 87.9%) than for the incidental learning condition (day one: 83.9%; day two: 83.1%). However, accuracy appeared slightly lower on day two than day one for the intentional learning condition, but remained at a similar level on both days for the incidental learning condition. The analysis revealed that this time the main effect of learning condition was non-significant [$\chi^2(1) = 3.66, p = .056$]. The main effect of time was also non-significant [$\chi^2(1) = 3.81, p = .051$], as was the main effect of position of the training task [$\chi^2(1) = 0.002, p = .966$]. Interestingly, the interaction between learning condition and time was significant [$\chi^2(1) = 3.85, p = .0497$]. The interaction between learning condition and training task position was non-significant [$\chi^2(1) = 2.24, p = .135$], as was the interaction between time and training task position [$\chi^2(1) = 0.008, p = .929$], and the three-way interaction [$\chi^2(1) = 0.48, p = .488$].

To follow up on the significant interaction between learning condition and time, two simple effects analyses were carried out to determine the significance of time within each of the two learning conditions separately. For the incidental learning condition there was no significant effect of time [$\chi^2(1) = 0.10$, $p = .750$], indicating no forgetting between day one and day two, however there was a significant effect of time for the intentional learning condition [$\chi^2(1) = 6.07$, $p = .014$]. (The p-values for these simple effects analyses were compared against a Bonferroni-corrected α of .025.)
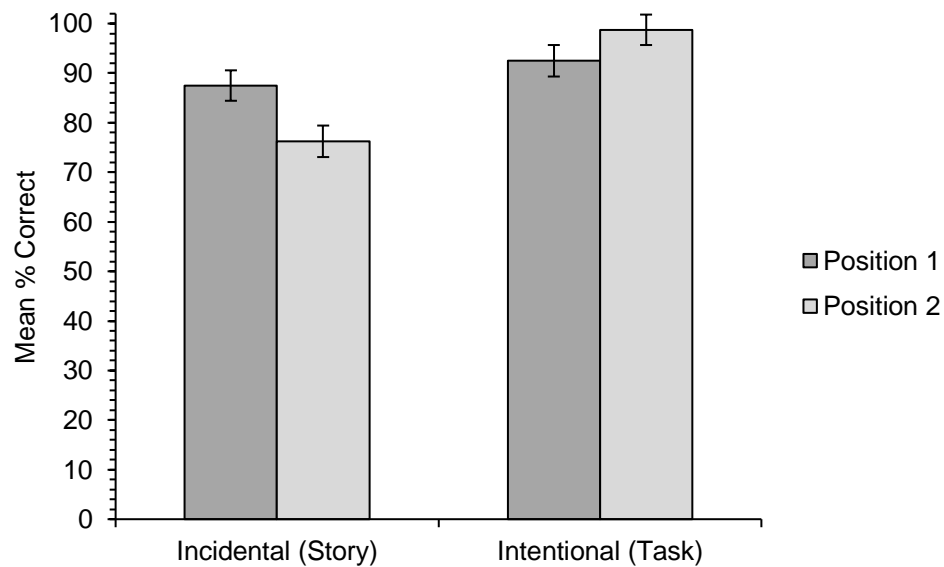
Figure 8. Experiment 2. Mean percentage of correct responses by subjects on the multiple-choice test (words correctly matched with the appropriate meaning) for each learning condition, when that condition was presented in the first or second position in the experiment. The data presented are for all participants only for the session on day one immediately after training ($N = 40$). Error bars show standard errors for subject means, adjusted for the within-participant factor of learning condition (Cousineau, 2005).

The mean percentage accuracy data for the multiple-choice test for all participants on day one only ($N = 40$) are shown in Figure 8. Accuracy was again high overall for both conditions, and slightly higher for items learned through the intentional condition (position one: 92.5%; position two: 98.8%) than for items learned through the incidental condition (position one: 87.5%; position two: 76.3%). There appeared to be no difference overall between items learned in the first or second position of the experiment, although the pattern of the means suggested an interaction in the same direction as for the cued recall accuracy data for day one only. However, the analysis showed that there was no significant main effect of

learning condition [$\chi^2(1) = 1.18$, $p = .277$], no significant main effect of task position [$\chi^2(1) = 0.004$, $p = .950$], and no significant interaction between these two variables [$\chi^2(1) = 2.52$, $p = .113$].

3.2.4 Discussion

This study aimed to determine how easily novel word meanings can be acquired incidentally through story reading, as compared with a more intentional learning procedure. This was investigated by testing participants' explicit memory for novel meanings assigned to previously unambiguous words, which were trained through story reading to allow for incidental learning and a repetitive training task designed to encourage intentional learning strategies. Memory for the new meanings was tested using both cued recall and a multiple-choice meaning-to-word matching recognition test immediately following training, and again after a 24-hour overnight delay. It was predicted that both tasks would facilitate adequate learning of the novel meanings, although no predictions were made as to which would produce greater learning, or if one training method would allow for better retention of the novel meanings over time.

The results showed that accuracy in recalling the new word meanings was significantly higher in the intentional learning condition than the incidental learning condition, accuracy was 85.2% for items learned under intentional conditions and 62.1% for items learned under incidental conditions when measured immediately after training. The accuracy data for the multiple-choice meaning-to-word matching test showed a similar pattern (96.0% for the intentional condition and 83.9% for the incidental condition at the immediate test), although for this measure the main effect of learning condition was non-significant. Furthermore, reading the story took participants a significantly longer amount of time, so more was learned in a shorter amount of time through the intentional training task. These findings are somewhat consistent with those of some L2 vocabulary learning studies that have found intentional learning to be more efficient than incidental learning of vocabulary (Hulstijn, 1992; Peters et al., 2009). Although recall accuracy was higher for the intentional learning condition, there was also a reasonably high level of acquisition of new meanings for familiar words through the incidental learning condition. This is very similar to the results of Experiment 1 of this thesis, where accuracy in recalling new meanings for familiar words was 63.5% after eight exposures in an incidental learning context (accuracy in cued recall of word forms was 69.2%). These results therefore support the findings of recent studies showing good acquisition of L1 vocabulary from reading (Batterink & Neville, 2011; Godfroid et al., 2017; Pellicer-Sánchez, 2016).

Furthermore, the present study replicates the finding from Experiment 1 of no significant forgetting of the new word meanings learned incidentally from stories between the immediate test and the delayed test. There was no significant main effect of time of test for the cued recall measure, for the meaning-to-word matching test it was marginally non-significant. The finding of no significant forgetting after 24 hours in the present experiment and after seven days in Experiment 1 is intriguing. A possible explanation for the lack of forgetting shown at the delayed test is that the additional retrieval practice for the test immediately after training may aid learning. Further supporting this possibility, items learned through the incidental condition showed a slight improvement (8% increase) in cued recall accuracy between the test on day one and the test on day two. The second test in the immediate test session (meaning-to-word matching) may therefore have boosted learning, manifesting as improved cued recall at the delayed test. The potential involvement of a testing effect (Roediger & Karpicke, 2006a) in long-term retention of new meanings for familiar words will be explored in detail in Experiment 3 and Experiment 4.

Interestingly, the recognition test showed a significant interaction between learning condition and time of test. Following up on this interaction, simple effects analyses showed that there was significant forgetting of items learned in the intentional condition (8% reduction in accuracy between the immediate test and delayed test), but there was no forgetting of items learned incidentally through story reading. A possible explanation for this finding is that new word meanings learned in a more semantically rich context, such as from stories, may be retained better. New word meanings encountered in stories will contain additional contextual information relating to the narrative (e.g., characters' thoughts and feelings), providing additional cues for participants to rely on for later retrieval.

The factor for position of each of the learning conditions in the experiment (first or second) was added to the models for the analysis to see whether performance in the incidental learning condition was better when it followed the intentional learning condition. It was feasible that participants could deduce the word-learning purpose of the study when carrying out the intentional learning task, which may cause participants to use more deliberate learning strategies upon encountering the new word meanings in the following story. The results for the cued recall measure showed a significant interaction between position of the training task in the experiment and learning condition only in the analysis for all participants who completed the test on day one. However, this interaction is in the opposite direction: items were recalled better from the story condition when it came first in the experiment, before the intentional learning condition. A possible explanation is that participants were concentrating better during the first task in the experiment, and may have become slightly fatigued by the time they got on to the second task, which may have had more of an impact on story reading

as it was the longer of the two tasks. Therefore, there was no concern that participants used the intentional training task as a cue to use deliberate learning strategies for the new meanings for familiar words in the stories.

In sum, the results from Experiment 2 show that acquiring new meanings for familiar words seems to be more efficient under intentional learning conditions than incidental learning conditions. However, new word meanings can also be learned reasonably well incidentally through story reading. As for Experiment 1, there was no significant forgetting of word meanings between the immediate and delayed test, in this case after a 24-hour delay. This was possibly due to retrieval practice at the immediate test having a beneficial effect on performance on the delayed test the following day; Experiment 3 and Experiment 4 will explore this possibility. Finally, the results of the recognition test suggest that new word meanings learned incidentally though stories may be retained better than those learned under intentional conditions, possibly due to the benefit of a more semantically rich learning context, although further research is required to investigate this.

3.3 Experiment 3: The testing effect in incidental and intentional learning

3.3.1 Introduction

Experiment 3 was preregistered through the Open Science Framework; the preregistration can be retrieved from https://osf.io/e5zmk (Hulme & Rodd, 2016, November 4). Where applicable any deviations from the preregistration are noted in the Method and Results sections for this experiment.

In the previous experiments in this thesis, participants showed very little evidence of forgetting at a surprise delayed test one day (Experiment 2) or even one week (Experiment 1) after training, when these new meanings were learned incidentally or intentionally (Experiment 2), and even with very few exposures to the word with its new meaning (Experiment 1). Previous research (for a review see Roediger & Karpicke, 2006b) has shown that additional retrieval practice at a memory test immediately following training can facilitate future long-term retention of learned information such as lists of FL vocabulary and their translations (Van den Broek et al., 2013), or information from prose passages (Butler, 2010). The testing effect has not previously been compared between incidental and intentional learning, and has not previously been examined for the learning of new meanings for familiar word forms.

The aim of Experiment 3 was therefore to investigate whether testing immediately after training improves long-term retention of novel meanings for familiar word forms after one day, and if this effect differs for novel meanings learned through incidental and intentional means. In a similar design to Experiment 2, participants were trained on new meanings for familiar words incidentally through reading one of the stories, as well as through the intentional learning condition. However, this time participants were tested immediately on only half of the items they saw in each learning condition; they were tested on all trained items at the surprise delayed test 24 hours later in which they were tested on half on the items for the first time, and retested on items that had been tested the previous day. As for Experiment 2, both test sessions consisted of cued recall of the new meanings, followed by a multiple-choice meaning-to-word matching recognition test. It was predicted that there would be better long-term retention for items that were tested immediately after training than those that were not. It was further predicted that the presence of an immediate test would enhance learning of items trained through both incidental and intentional means. There was no prediction as to whether the magnitude of the testing effect would differ for the different learning conditions. Additionally, it was hypothesised that retention would be better overall for novel meanings learned through intentional means than for those learned under incidental conditions, but items learned incidentally may be retained better over time as seen in Experiment 2.

3.3.2 Method

Participants

Ninety-nine adult participants took part and were included in the experiment (age: $M = 32.31$ years, $SD = 8.14$, range = 18-49), 56 of whom were female. All participants were monolingual native British English speakers and had not been diagnosed with any reading or language impairments. Participants were recruited through the website Prolific Academic (Damer & Bradley, 2014), and were paid for their participation at the end of each session (£6 for session one and £2 for session two).

In addition to the 99 participants included in the study, an additional 36 participants took part in the first session of the experiment, but did not return to complete session two within the deadline and were excluded from the study. A further twenty-one participants were excluded due to getting more than one of the multiple choice comprehension questions wrong in the story-reading training condition, and two further participants were excluded for attempting to do the experiment more than once. Finally, five participants were excluded for being outliers in their mean reading speeds (faster than 543.4 words per minute, 2 SD above

the mean). The excluded participants were replaced with new participants to obtain the total of 99 participants included in the study.

Materials

The stimuli for the present experiment were identical to those used in the previous experiments in this thesis. These were the 16 known words with novel semantically unrelated noun meanings (see Appendix A), which had been incorporated into the four separate short stories that were used in the incidental learning condition (see Appendix B). The same sentences that had been used in Experiment 2 were also used in the intentional learning condition in this experiment (see Appendix A), as well as the short paraphrased versions of the definitions (see Appendix C). The same longer paraphrased versions of the definitions were again used in the multiple-choice meaning-to-word matching test task (see Appendix D).

Design

The experiment used a three by two within-subjects design, with two independent variables: learning condition (2 levels: incidental versus intentional) and test type (3 levels: immediate test (tested in the first session) versus delayed test (tested for the first time in the second session) versus delayed retest (tested for the second time in the second session)). The dependent variables were accuracy on the cued recall and multiple-choice meaning-to-word matching tests.

There were sixteen versions of the experiment in total, this was to ensure that all of the stimulus items were seen an even number of times in each condition, with the order of the conditions counterbalanced across participants. As in Experiment 2, each participant was trained on half the total number of stimuli (eight items), as this was deemed to be a reasonable number of new word meanings to learn in a single session. Half of the participants were therefore trained on the set of items that appeared in Stories 1 and 4, and the other half were trained on the set of items that appeared in Stories 2 and 3. Within the first set of items, the words occurring in Story 1 were presented in the incidental learning condition for half the participants, with the words from Story 4 being trained through the intentional learning condition, and vice versa, with the same organisation for the second set of items. Additionally, the order of the incidental learning condition and intentional learning condition was counterbalanced across all participants to minimise any order effects of the different tasks. Finally, the stimulus items that were or were not tested immediately following training were

also counterbalanced across participants by dividing the items in each of the two sets of items into two 'testing sets'. The two 'testing sets' for the first set of items each contained two items which appeared in Story 1 and two items which appeared in Story 4, and the items were grouped such that the potential difficulty of items in the two sets was balanced (based on item performance from the previous experiments), with the same arrangement for the second set of items. Participants were randomly assigned to one of the sixteen versions of the experiment.

Procedure

The first session of the experiment was almost exactly the same as Experiment 2. The procedure for the incidental and intentional learning conditions was identical to that of Experiment 2, and immediately following training participants completed the same Mill Hill vocabulary test (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996; Raven et al., 1998) as a filler task. Participants were then given an immediate test on half of the items that they had been trained on through the incidental and intentional learning conditions (four items, two trained through each training method). The test tasks used were the same as for Experiment 2: cued recall followed by meaning-to-word matching. The stimulus items were tested in a random order in each of the two tasks, with no feedback given to participants. In the meaning-to-word matching test task, only the four stimulus words that a participant was being immediately tested on appeared as the four alternative responses to choose from for each test item; the order of these was also randomised for each test item.

Exactly 24 hours after the first session of the experiment had been made available to participants, the participants were asked to take part in the second session of the experiment: the delayed test. The participants were not aware beforehand that they would be asked to complete this test in order to discourage them from rehearsing and intentionally retaining information about the novel word meanings. As such, unfortunately 36 participants did not return to complete session two. The participants started the delayed test an average of 24 hours and 25 minutes (SD = 57 mins, range = 22h45m-27h21m) after they had begun the training session. The test tasks used for the delayed test were the same as those that had been used for the immediate test: cued recall followed by meaning-to-word matching. This time participants were tested on all of the stimuli that they had been trained on (eight items). The order of presentation of the items in each of the two tests was again randomised for each participant, and for the meaning-to-word matching test the order of the eight stimulus words to choose from was again randomised for each test item.

3.3.3 Results

Analysis procedure

As with the previous experiments, responses for the cued recall test were coded by the experimenter blind to condition using the same procedure described in Experiment 1. Responses were coded into the binary accuracy data ("1" for correct and "0" for incorrect) for analysis. As in Experiment 2, responses for the multiple-choice test were also simply coded as "1" for correct or "0" for incorrect depending on whether the appropriate word had been selected to match with the meaning presented on a given trial.

Upon completion of the experiment, it was noted that test type was confounded by a difference in test difficulty between the immediate test and the two delayed test types for the multiple-choice meaning-to-word matching measure. This was because in the immediate test participants had to choose from among four alternative words to pair with the appropriate new meaning on each trial, while in the delayed test participants had to choose from among eight alternative words. The results from the immediate multiple-choice test are therefore not comparable to the results from the two delayed test types, and so the analysis for this measure was only carried out on the subset of results for the two delayed test types. This is a deviation from the analysis plan outlined in the preregistration of this experiment; the analysis of the cued recall measure was carried out according to the preregistration.

The data from both the cued recall test and multiple-choice meaning-to-word matching test were analysed using logistic mixed effects models using the 'lme4' package, (version 1.1-12; Bates et al., 2015) and R statistical software (version 3.3.2, R Core Team, 2017). Separate models were fitted for the analysis of the data from the two measures. The model used in the analysis of the cued recall data contained three factors: test type (3 levels: immediate, delayed 1st, delayed 2nd), learning condition (2 levels: incidental, intentional), and position of the training task in the experiment (2 levels: 1st, 2nd). The contrasts for the fixed effect of test type were defined using Helmert coding, with one contrast comparing the immediate test to the two delayed tests combined (immediate: 0.67, delayed 1st: -0.33, delayed 2nd: -0.33), and a second comparing the two delayed test types to each other (immediate: 0, delayed 1st: -0.5, delayed 2nd: 0.5). Deviation coding was used to specify the contrasts for the fixed effects of learning condition (incidental: -0.5, intentional: 0.5) and task position (1st: -0.5, 2nd: 0.5).

The model used for the analysis of the multiple-choice data had also three factors: test type (2 levels: delayed 1st, delayed 2nd), learning condition (2 levels: incidental, intentional), and task position (2 levels: 1st, 2nd). The contrasts were specified using deviation coding for

the fixed effects of test type (delayed 1st: -0.5, delayed 2nd: 0.5), learning condition (incidental: -0.5, intentional: 0.5), and task position (1st: -0.5, 2nd: 0.5).

As for the previous experiments, recommendations by Barr et al. (2013) were followed for selection of the appropriate random effects structure for the models. The first attempted fit used the maximal random effects structure[11], with a random intercept and slopes for test type, learning condition, and the interaction by participants[12]; and a random intercept and slopes for test type, learning condition, and the interaction by items. This model converged for the multiple-choice meaning-to-word matching measure, and was thus used as the final model for the analysis; the model for the cued recall test did not converge. Following this, the model for the cued recall measure was simplified by removing the correlations between the random slopes and intercepts (without removing any of the random slopes), which allowed the model to converge, so this was used as the final model for the analysis.

Significance of the fixed effects and interactions were assessed using likelihood ratio tests comparing the full model to models with each of the factors/interactions of interest removed in turn (but leaving in any of the other interactions involving that factor/interaction) and leaving the random effects structure intact.

Following on from the main analysis for the cued recall measure, firstly three pairwise comparisons (with Bonferroni adjustment for multiple comparisons, $\alpha = .017$) were made between the different levels of test type to determine which test types were significantly different from each other. These pairwise comparisons were run by taking a subset of the data for each pair of levels of test type and creating a model for each containing the same fixed and random effects as the model used for the main analysis, although the contrast for test type was coded using deviation coding in each model (immediate: 0.5 vs. delayed 1st test: -0.5; immediate: 0.5 vs. delayed 2nd test: -0.5; delayed 1st: -0.5 vs. delayed 2nd test: 0.5). Significance of the effect of test type for each of the pairwise analyses was determined using likelihood ratio tests (comparing the model containing the factor of interest to an identical model with that factor removed). Secondly, for the cued recall measure follow-up pairwise comparisons (with Bonferroni correction for multiple comparisons, $\alpha = .017$) were made for the three 2x2 interactions between the pairs of test types and the two learning conditions to determine whether the difference between any two test types was significantly different between the two

---

[11] The "bobyqa" optimiser was used as per recommendations by Bates et al. (2016) for dealing with model convergence issues.

[12] Random slopes for task position by participants and by items were not included due to issues with model non-convergence.

learning conditions. This was done using the same models as for the previous follow-up analysis and using likelihood ratio tests to compare each of those models to an identical one with the interaction between learning condition and test type removed.

Finally, six simple effects subset pairwise comparisons (with Bonferroni adjustment for multiple comparisons, $\alpha = .008$) were run for the cued recall measure to test for any significant differences between the different test types within the two learning conditions. This was done by taking further subsets of the data for the pairs of levels of test type separately for the incidental and intentional learning conditions and creating models with only fixed effects for test type, task position, and the interaction (and random effects for test type by participants and items). Significance for the simple effects was again assessed using likelihood ratio tests comparing each of the models to an identical one with only the fixed effect for test type removed in each instance.

The only follow-up analyses carried out for the multiple-choice meaning-to-word matching test were two simple effects subset pairwise comparisons (with Bonferroni correction for multiple comparisons, $\alpha = .025$). This was carried out in the same way as the simple effects analyses were carried out for the cued recall measure.

Cued recall of novel meanings

The mean percentage accuracy data for the cued recall test (Figure 9) showed that accuracy was reasonable high overall, and higher for items trained through the intentional than the incidental learning condition. Cued recall accuracy appeared higher in the immediate test and items tested for the second time at the delayed test than for items tested for the first time at the delayed test. There also appeared to be an interaction between learning condition and test type whereby cued recall performance for intentionally-trained items appeared to remain the same between the immediate and delayed 2nd tests, while for incidentally-trained items there was an improvement in accuracy between the immediate test and when these items were retested after the delay.
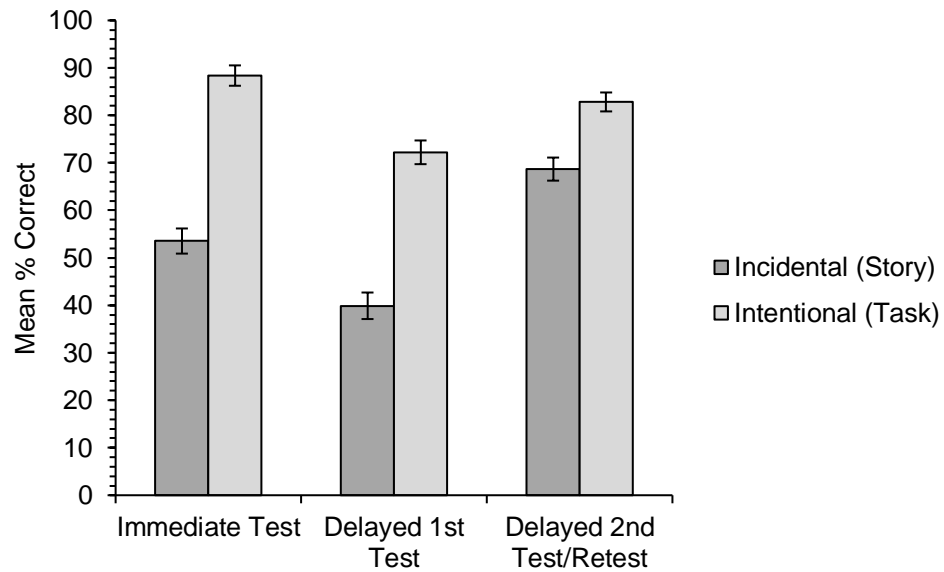
Figure 9. Experiment 3. Mean percentage of correct responses given by participants in the cued recall test[13] (meanings correctly recalled for the appropriate word) for each learning condition and for the three different test types in the experiment. Error bars show standard error of the subject means adjusted for the within-participant design (Cousineau, 2005).

The cued recall accuracy data (see Figure 10) also showed that there did not appear to be a great difference in performance between items that were trained in the first or second position in the experiment. Importantly, there also did not appear to be an interaction between learning condition and task position; cued recall accuracy remained similar for items trained in the first or second position in the experiment through both training methods.

---

[13] The LME analyses were carried out on the raw binary accuracy data, however mean percentage accuracy data are displayed in the graphs for ease of interpretation.
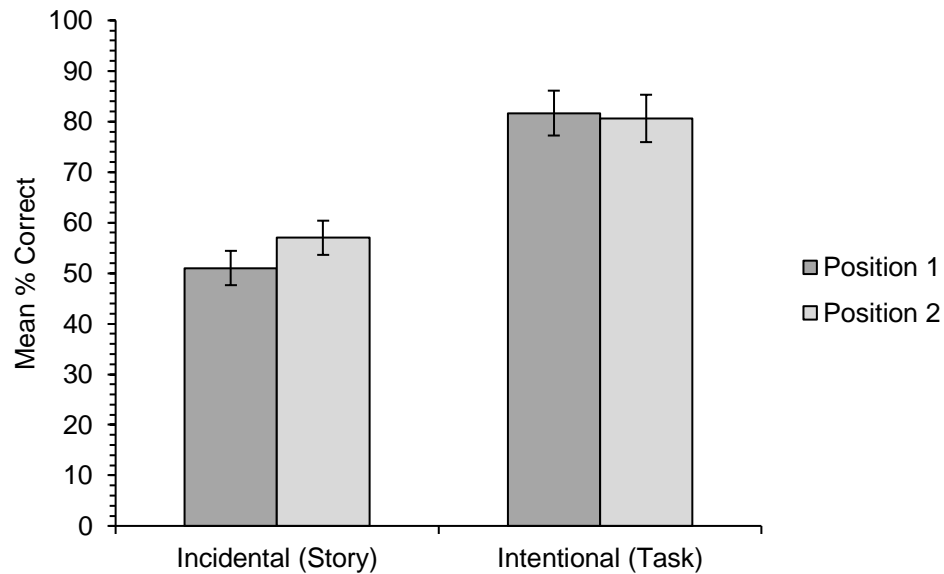
Figure 10. Experiment 3. Mean percentage of correct responses given by participants in the cued recall test (meanings correctly recalled for the appropriate word) for each learning condition, when that condition was presented in the first or second position in the experiment. Error bars show standard error of the subject means adjusted for the within-participant design (Cousineau, 2005).

The main effect of learning condition was significant [$\chi^2(1) = 34.83$, $p < .001$], with more meanings trained through the intentional task being correctly recalled than those trained incidentally through the stories. The main effect of test type was also significant [$\chi^2(2) = 25.78$, $p < .001$], and the main effect of task position was non-significant [$\chi^2(1) = 2.50$, $p = .114$]. There was a significant interaction between learning condition and test type [$\chi^2(2) = 13.86$, $p < .001$], but the interaction between learning condition and task position was not significant [$\chi^2(1) = 0.09$, $p = .760$]. There was an unexpected significant interaction between test type and task position [$\chi^2(2) = 9.24$, $p = .010$] (see Figure 11); the three-way interaction was not significant [$\chi^2(2) = 3.23$, $p = .199$].
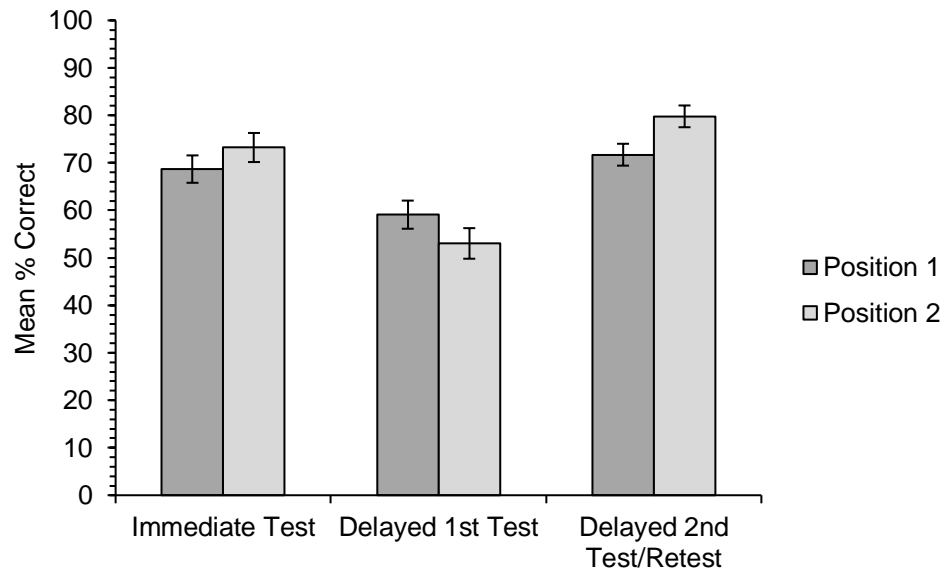
Figure 11. Experiment 3. Mean percentage of correct responses given by participants in the cued recall test (meanings correctly recalled for the appropriate word) for the three different test types in the experiment, when items were learned in the task presented in the first or second position in the experiment (averaged across learning conditions). Error bars show standard error of the subject means adjusted for the within-participant design (Cousineau, 2005).

To further investigate the significant main effect of test type, three pairwise comparisons between the different levels of test type were carried out. The results revealed that there was a significant effect of overnight forgetting (difference between the delayed 1st and immediate tests) $[\chi^2(1) = 19.99, p < .001]$, with better recall of new meanings when tested immediately than when tested for the first time after a delay. There was also a significant testing effect (difference between the delayed 2nd and delayed 1st tests) $[\chi^2(1) = 18.83, p < .001]$, with higher recall accuracy for items that were being tested for the second time than for those being tested for the first time after the delay. However, there was no significant difference in cued recall accuracy between the immediate and delayed 2nd tests $[\chi^2(1) = 0.89, p = .345]$. (The p-values for these comparisons were compared against a Bonferroni-corrected α of .017).

To further investigate the significant interaction between learning condition and test type, the second level of follow-up analyses were pairwise comparisons for the three 2x2 interactions between the pairs of test types and the two learning conditions. The results showed that there was a significant interaction between learning condition and the difference between the immediate and delayed 2nd tests $[\chi^2(1) = 16.24, p < .001]$. Items learned incidentally from stories showed some improvement between the immediate test and the retest on day two, while

items learned through the intentional learning condition showed a small amount of forgetting. There was no significant interaction between learning condition and the difference between either the immediate and delayed 1st tests [$\chi^2(1) = 2.29$, $p = .130$] or between the delayed 1st and delayed 2nd tests [$\chi^2(1) = 2.97$, $p = .085$]. (The p-values for these comparisons were compared against a Bonferroni-corrected α of .017).

In the final level of follow-up analyses for the cued recall measure, six simple effects subset pairwise comparisons were run to test for any significant differences between the different test types within the two learning conditions. (The p-values for these comparisons were compared against a Bonferroni-corrected α of .008). The results revealed that, for the incidental learning condition there was no significant difference in recall accuracy for items tested for the first time after the delay than for items tested immediately after training at the corrected level [$\chi^2(1) = 5.88$, $p = .015$]. There was significantly better cued recall accuracy for items tested for the second time after the delay than for those tested for the first time [$\chi^2(1) = 15.27$, $p < .001$], and for the incidental learning condition there was also significantly better recall of new meanings tested for the second time after the delay than the immediate test [$\chi^2(1) = 14.59$, $p < .001$]. For the intentional learning condition, there was significantly lower recall accuracy for items tested for the first time after the delay than those tested immediately [$\chi^2(1) = 18.27$, $p < .001$], and there was again significantly better recall accuracy for items tested for the second time after the delay than those tested for the first time [$\chi^2(1) = 8.39$, $p = .004$]. However, for the intentional learning condition there was no significant difference (at the corrected level) in cued recall accuracy between items tested for the second time following the delay and when tested immediately after training [$\chi^2(1) = 4.25$, $p = .039$].

Additionally, although not pre-specified in the preregistration for this experiment, exploratory follow-up analyses were carried out to examine the nature of the unexpected interaction between test type and position of the learning task in the experiment. Three pairwise comparisons were made of the 2x2 interactions between the pairs of test types and the two training task positions (first or second in the experiment). The results revealed a significant interaction between position and the difference between the delayed 1st and delayed 2nd tests [$\chi^2(1) = 7.12$, $p = .008$]. Items appeared to be recalled better at the delayed 1st test when they had been presented in the first position in the training session, whereas items were recalled better at the delayed 2nd test when they had been trained in the second part of the training session. There was no significant interaction between the immediate and delayed 2nd tests [$\chi^2(1) = 0.12$, $p = .729$], nor between the immediate and delayed 1st tests [$\chi^2(1) = 5.15$, $p = .023$] at the Bonferroni-corrected level (α = .017). Although, further follow-up analyses of the simple effects pairwise comparisons between the two training positions within the delayed

1st and delayed 2nd test types were both non-significant at the Bonferroni-corrected level (both $p > .025$).

Multiple choice meaning-to-word matching

The mean percentage accuracy data for the multiple-choice meaning-to-word matching test (Figure 12) showed that overall accuracy was very high, and accuracy was higher for items trained intentionally than incidentally across all test types. For the delayed test types it appeared that meanings were correctly paired with the appropriate word more accurately for items that were being tested for the second time than those that were being tested for the first time after the 24-hour delay. There also appeared to be an interaction between learning condition and test type, with a similar level of accuracy for intentionally-trained items regardless of whether or not they had been tested the previous day, while for incidentally-trained items there is higher accuracy for items that had been previously tested than those that had not.
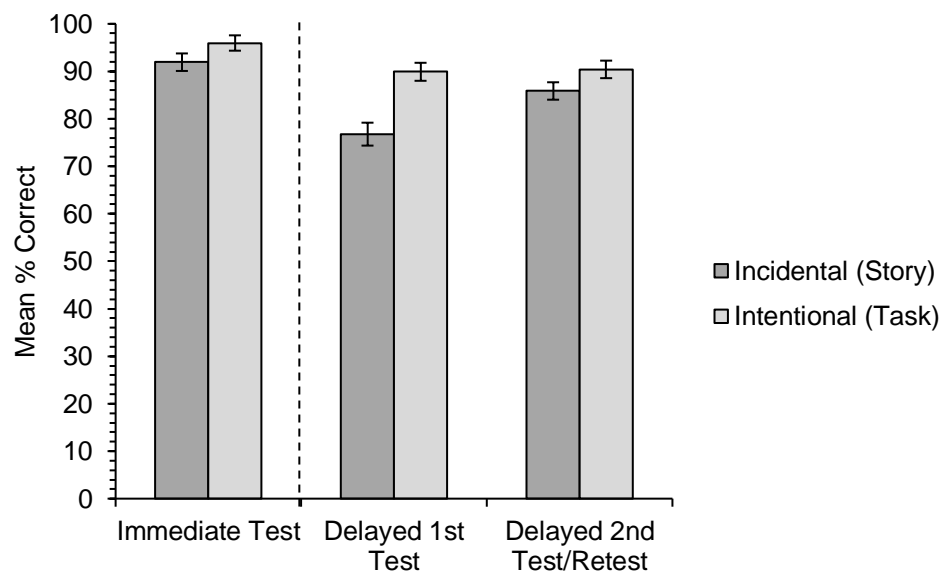
Figure 12. Experiment 3. Mean percentage of correct responses given by participants in the multiple-choice test (meanings correctly matched with the appropriate word) for each learning condition and for the three different test types in the experiment. Note that the results from the immediate test are not comparable to those from the two delayed test types due to an underlying difference in test difficulty. Error bars show standard error of the subject means adjusted for the within-participant design (Cousineau, 2005).

As was the case for the cued recall data, the data for the multiple-choice meaning-to-word matching test averaged across test type (Figure 13) appeared to show no great difference in mean percentage accuracy between items trained in the first and second positions in the experiment. Again, there did not appear to be an interaction between learning condition and task position, with similar levels of accuracy for items trained in the first and second positions in the experiment through both training methods.
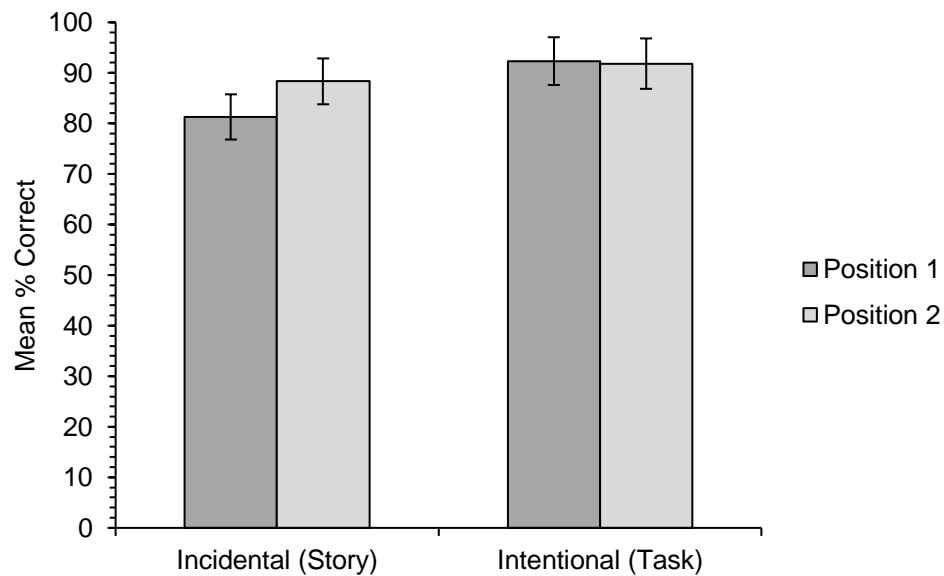
Figure 13. Experiment 3. Mean percentage of correct responses given by participants in the multiple-choice test (meanings correctly matched with the appropriate word) for each learning condition, when that condition was presented in the first or second position in the experiment. Error bars show standard error of the subject means adjusted for the within-participant design (Cousineau, 2005).

The main effect of learning condition was significant [$\chi^2(1) = 8.44$, $p = .004$], with higher accuracy again for items trained intentionally than incidentally. The main effect of test type was also significant [$\chi^2(1) = 6.71$, $p = .010$], with slightly greater accuracy for items that had been tested previously than for those that had not been; there was no significant main effect of task position [$\chi^2(1) = 0.32$, $p = .569$]. The interaction between learning condition and test type was not significant [$\chi^2(1) = 0.61$, $p = .435$], nor was the interaction between learning condition and task position [$\chi^2(1) = 1.23$, $p = .268$], nor the interaction between test type and task position [$\chi^2(1) = 1.32$, $p = .251$]. The three-way interaction was also not significant [$\chi^2(1) = 0.06$, $p = .810$].

Following on from the main analysis, two simple effects subset pairwise comparisons tested for any significant differences between the different test types within the two learning conditions. The results showed that for the incidental learning condition, the difference between items tested for the second time following the delay and those tested for the first time was non-significant at the corrected level [$\chi^2(1) = 4.62$, $p = .032$]. For the intentional learning condition, there was no significant difference in accuracy between items tested for the first or second time after the delay [$\chi^2(1) = 3.53$, $p = .060$]. (The p-values for these comparisons were compared against a Bonferroni-corrected α of .025).

3.3.4 Discussion

The aim of Experiment 3 was to examine whether testing memory immediately after training enhances long-term retention of new meanings for familiar words acquired through incidental and intentional learning conditions, and to see whether any testing effect differs depending on the learning conditions. The method used in Experiment 2 was adapted for the present study. New meanings for familiar words were trained under both learning conditions, and participants' memory for only half of the items they had been trained on was tested immediately through cued recall and a multiple-choice meaning-to-word matching recognition test; memory for all new meanings was then tested 24 hours later. It was predicted that items trained through both learning conditions would be remembered better after 24 hours if they had been tested immediately following training, but there was no prediction as to whether this effect would differ between the two learning conditions.

The present experiment replicated the finding of Experiment 2 that new meanings for familiar words were learned better overall through intentional learning conditions than through incidental learning conditions. There was a significant main effect of learning condition for both measures, so overall accuracy was higher for items trained through the intentional condition for both cued recall and recognition. This finding will be explored in greater detail alongside the findings from Experiment 2 in the general discussion of this chapter.

Cued recall accuracy was higher overall immediately after training than when items were tested for the first time after 24 hours. This demonstrates some overnight forgetting of the new meanings for the words in the absence of an intervening test. Furthermore, there was numerically but not significantly (at the corrected level) more forgetting of items trained through the intentional learning condition than those learned through a story, both with and without prior retrieval practice. This is in line with the findings of Experiment 2, where

recognition accuracy was lower after 24 hours for items learned through the intentional training task, but not for items learned through the stories.

Both the cued recall and meaning-to-word matching tests revealed an overall testing effect: new meanings for familiar words were recalled and recognised significantly better after 24 hours when they had been tested immediately after training than when they were being tested for the first time on day two, with no feedback on performance at any time. As predicted, in the cued recall measure this main effect of testing was also significant in the simple effects that looked at incidental and intentional learning separately. This result is in line with studies that have found a benefit of prior retrieval on learning information from different contexts, such as list of FL vocabulary words and their translations (Van den Broek et al., 2013) and information from prose passages (Roediger & Karpicke, 2006b).

The lack of difference in cued recall accuracy between performance on the immediate test and performance on the delayed test for items being tested for the second time shows that the retrieval practice protected these items against forgetting. The testing effect seen in the present study therefore at least partly explains why participants in Experiment 2 and Experiment 1 showed such good retention after one day and one week respectively. In the present study both the cued recall and recognition tests were administered to all participants at both time points. It is therefore unclear whether either of these tests on its own would produce a testing effect, or if it was the combination of the two that was important for boosting long-term retention. It also remains to be seen whether one of these test types is better than the other for enhancing retention of new meanings for familiar words learned incidentally through reading. Experiment 4 will address this issue.

To summarise, Experiment 3 demonstrated that testing memory of new meanings for familiar words benefits their future retention. This was the case for recalling word meanings learned either incidentally through story reading or through an intentional learning condition. As in Experiment 2, participants learned vocabulary more efficiently through the intentional learning condition, but performance for both learning conditions was good. There was non-significantly less forgetting of items trained incidentally through the stories, and the testing effect was also non-significantly larger for incidentally-trained items which seemed to benefit from the additional learning opportunity afforded by the immediate test. Either the immediate cued recall or meaning-to-word matching test, or indeed a combination of the two, may have produced the observed testing effect, and Experiment 4 will explore which of these test methods could be more beneficial for future retention.

3.4 Experiment 4: Immediate test method and incidental learning

3.4.1 Introduction

Experiment 4 was also preregistered through the Open Science Framework; the preregistration is available at https://osf.io/c59tz (Hulme & Rodd, 2017, June 23). Again, where applicable any deviations from the preregistration are noted in the Method and Results sections for this experiment.

Experiment 3 of this thesis demonstrated that testing participants immediately after training on both their cued recall and recognition of new meanings for familiar word forms enhanced long-term retention of those new word meanings. This testing effect was present for new word meanings learned under both incidental and intentional learning conditions, and was non-significantly larger for new meanings for familiar words learned incidentally from stories. However, it is unclear whether the immediate test of cued recall or recognition alone would elicit the same effect, or whether one of these two test methods is more beneficial than the other for retention of new meanings for familiar words learned through reading.

The aim of Experiment 4 was therefore to investigate the impact of the method of immediate test (cued recall compared with meaning-to-word matching) on the testing effect in the long-term retention of novel meanings for familiar words learned incidentally through story reading. In this experiment participants learned new meanings for familiar words incidentally through reading two stories. Their memory was then tested immediately on half of the items trained in each story either through a test of cued recall or recognition (multiple-choice meaning-to-word matching), with long-term retention assessed at a delayed test 24 hours later using both test measures (cued recall followed by multiple-choice meaning-to-word matching).

There were three possible outcomes for Experiment 4. The first was that cued recall would be more beneficial for long-term retention as, according to the retrieval effort hypothesis (Pyc & Rawson, 2009), production tests that require more effortful retrieval than recognition tests (Roediger & Butler, 2011) are more helpful for retention. The second possibility was that meaning-to-word matching would be more useful for future retention as stronger cues are provided which allow for an additional learning opportunity (Marsh et al., 2007). The third possibility was that the testing effect would not transfer across test tasks (Hogan & Kintsch, 1971; Tran et al., 2014), and so the benefit of each method of immediate

test would only be seen for the delayed test of the same type, in which case it could be characterised as more of a practice effect.

It was predicted that, as for Experiment 3, there would be better long-term retention for items tested immediately after training through either method than those that were not tested previously. Additionally, based on the findings of previous studies (e.g., McDaniel et al., 2007) it was hypothesised that cued recall would be more beneficial for long-term retention of new meanings for familiar words than multiple-choice meaning-to-word matching, as tests that require greater retrieval effort may be more useful for learning.

3.4.2 Method

Participants

Ninety-eight adult participants took part and were included in the experiment (age: $M = 33.7$ years, $SD = 8.0$, range = 18-49), 64 of whom were female. All participants were monolingual native British English speakers and had not been diagnosed with any reading or language impairments. Participants were recruited through the website Prolific Academic (Damer & Bradley, 2014), and were paid for their participation at the end of each session (£4 for session one and £2 for session two).

In addition to the 98 participants included in the study, an additional 18 participants took part in the first session of the experiment, but did not return to complete session two within the deadline and were therefore excluded from the study. A further thirty-five participants were excluded due to getting more than one of the multiple choice comprehension questions wrong in either of the stories they read. Seven further participants were excluded due to a technical issue during data collection, and two participants were excluded for being outliers in their mean reading speeds (faster than 806.2 words per minute, 2 SD above the mean for all participants not already excluded for one of the aforementioned reasons ($M = 308.7$ words per minute)). The excluded participants were replaced with new participants to obtain the total of 98 participants included in the study.

Materials

The stimuli for the present experiment were identical to those used in the previous experiments in this thesis. These were the 16 real English words with novel semantically unrelated meanings (see Appendix A), which had been incorporated into the four separate short stories

(see Appendix B). The same paraphrased versions of the meaning definition sentences with the stimulus words removed that had been used in Experiment 2 and Experiment 3 were also used in the multiple-choice meaning-to-word matching test task in the present experiment. One additional different, paraphrased version of each of the sentences was created so that a differently worded definition would be presented in the immediate and delayed meaning-to-word matching tests in order to counteract any direct practice effects (see Appendix D).

Design

The experiment used a two by two mixed design, with two independent variables: immediate test method (two levels: cued recall versus meaning-to-word matching) manipulated between subjects, and the within-subjects variable of whether items were or were not previously tested (two levels: not pre-tested (tested for the first time in the second session) versus pre-tested (tested for the second time in the second session). The dependent variables were accuracy on the tests of cued recall of the new meanings and eight-alternative multiple-choice meaning-to-word matching measured at the delayed test time point.

There were eight versions of the experiment in total, this was to ensure that all of the stimulus items were seen an even number of times in each condition counterbalanced across participants. As in Experiments 2 and 3, each participant was trained on half the total number of stimuli (eight items per participant), as this was deemed to be a reasonable number of new word meanings to learn in a single session. Half of the participants were therefore trained on the set of items which appeared in Stories 1 and 4, and the other half were trained on the set of items which appeared in Stories 2 and 3. For the key factor of immediate test method, half of the participants ($N = 49$) had a cued recall test of half of their items (four items) immediately after training, and the other half of the participants ($N = 48$) had a multiple-choice meaning-to-word matching test of half of their items immediately after training. Finally, the stimulus items that were or were not tested immediately following training were also counterbalanced across participants by dividing the items in each of the two word sets into two 'testing sets'. The two 'testing sets' for the first set of items each contained two items which appeared in Story 1 and two items which appeared in Story 4, and the items were grouped such that the potential difficulty of items in the two sets was balanced (based on item performance from the previous experiments), with the same arrangement for the second set of items. Participants were randomly assigned to one of the eight versions of the experiment.

Procedure

The first session of the experiment began with the incidental training procedure. Participants first read one of the short stories with interleaved multiple choice comprehension questions after each page; the procedure for this was identical to that of the previous experiments. They were then asked to rate how enjoyable and clear they found the story, and answer some questions about their subjective reading style, which took around two minutes. Participants then read a second story with interleaved multiple choice comprehension questions after each page. Immediately following training participants completed the same Mill Hill vocabulary test as used in Experiments 2 and 3 (Mill Hill Vocabulary Test, Set A: Multiple Choice: Buckner et al., 1996; Raven et al., 1998) as a filler task. Participants were then given an immediate test of half on the items that they had been trained on (four items, two trained through each story), which was either a cued recall test or a multiple-choice meaning-to-word matching test depending on which version of the experiment they had been assigned to. The stimulus items were tested in a randomised order in both of the test tasks, with no feedback given to participants. In the multiple-choice meaning-to-word matching test task, only the four stimulus words that a participant was being immediately tested on appeared as the four alternative responses to choose from for each test item; the order of these was also randomised for each test item.

Exactly 24 hours after the first session of the experiment had been made available to participants, the participants were asked to take part in the second session of the experiment: the delayed test. The participants were not aware beforehand that they would be asked to complete this test in order to discourage them from rehearsing and intentionally retaining information about the novel word meanings, and to keep the procedure as similar as possible to that of Experiments 2 and 3. (As such, as mentioned previously, unfortunately 18 participants did not return to complete session two and were replaced during data collection.) The participants completed the delayed test an average of 24 hours and 31 minutes ($SD = 57$ mins; range = 22h40m-27h25m) after they had begun the training session the previous day. The test tasks used for the delayed test were the same as those that had been used for the immediate test, but this time participants completed both tests: cued recall followed by multiple-choice meaning-to-word matching. At the delayed test participants were tested on all of the stimuli that they had been trained on (eight items). The order of presentation of the items in each of the two tests was again randomised for each participant. For the meaning-to-word matching test, different paraphrased versions of the definition sentences were used to those that had appeared in the immediate test, and the order of the eight stimulus words to choose from was randomised for each test item.

3.4.3 Results

Analysis procedure

The responses on the cued recall and multiple-choice meaning-to-word matching tasks were coded in the same way as for the previous experiments. The data were analysed using linear mixed effects (LME) models with the lme4 package (version 1.1-13; Bates et al., 2015) and R statistical software (version 3.3.3; R Core Team, 2017). Two models were created to analyse the results of the two delayed tests of cued recall and eight-alternative multiple-choice meaning-to-word matching separately.

Both LME models contained random effects for participants and items (with a slope by participants for whether items were or were not previously tested; and slopes by items for whether items were or were not previously tested, the method of immediate test, and the interaction between these variables). The LME models for both tests also contained fixed effects for whether items were tested immediately after training (two levels: previously tested, or not previously tested), the method of the immediate test (two levels: cued recall, or meaning-to-word matching), and the interaction between these two variables (which was created by multiplying the contrasts for these two variables). The contrasts were defined using deviation coding for whether items were or were not immediately tested (not previously tested: -0.5 versus previously tested: 0.5), and the immediate test method (cued recall: -0.5 versus meaning-to-word matching: 0.5).

As for the previous experiments, the first attempted model fit in each case was with the maximal random effects structure (as recommended by Barr et al., 2013); the models for both test measures converged with the maximal random effects structure[14]. However, for the multiple-choice meaning-to-word measure a later model required to calculate significance of the main effect of whether items had been previously tested did not converge with the maximal random effects structure. The full model for this test measure therefore had to be simplified (following Barr et al., 2013) by removing the correlations between the random slopes and random intercepts for the random effects by participants and items (without removing any of the random slopes themselves). This allowed the full model for the multiple-choice meaning-to-word matching measure to converge.

---

[14] The "bobyqa" optimiser was used as per recommendations by Bates et al. (2016) for dealing with model convergence issues.

Significance of the main effects and interaction was assessed using likelihood ratio tests that compared the full models to identical models with only the factor or interaction of interest removed in turn (but leaving in any other interaction or main effect involving that factor or interaction, and always leaving the random effects structure intact).

Following on from the main analysis, planned simple effects analyses were carried out to determine the significance of the main effect for whether items had or had not been immediately tested within each of the two immediate test methods. This was done by taking a subset of the data for each of the two immediate test method groups and creating a model for each containing only a fixed effect for whether items were or were not immediately tested (as well as random effects, with a slope by participants and by items for whether items were or were not immediately tested). This was again carried out separately for the two delayed test measures (the model for the simple effects analysis for the multiple-choice meaning-to-word matching test had the random correlations removed to match with the full model used previously in the main analysis for this measure). Significance of the simple effects of whether items were or were not immediately tested within the two immediate test methods was determined using likelihood ratio tests comparing the models containing the factor of interest to one without (while retaining the random effects structure).

Cued recall of novel meanings

The mean percentage accuracy data for cued recall (measured in the delayed test session) are shown in Figure 14. Cued recall performance was low overall when items had not been tested immediately after training and was of a similar level for the immediate cued recall group (26.5%) and immediate multiple-choice meaning-to-word matching group (25.5%). Performance was much higher when items had been tested immediately after training, appearing higher for the immediate multiple-choice meaning-to-word matching group (58.3%) than for the immediate cued recall group (49.5%).
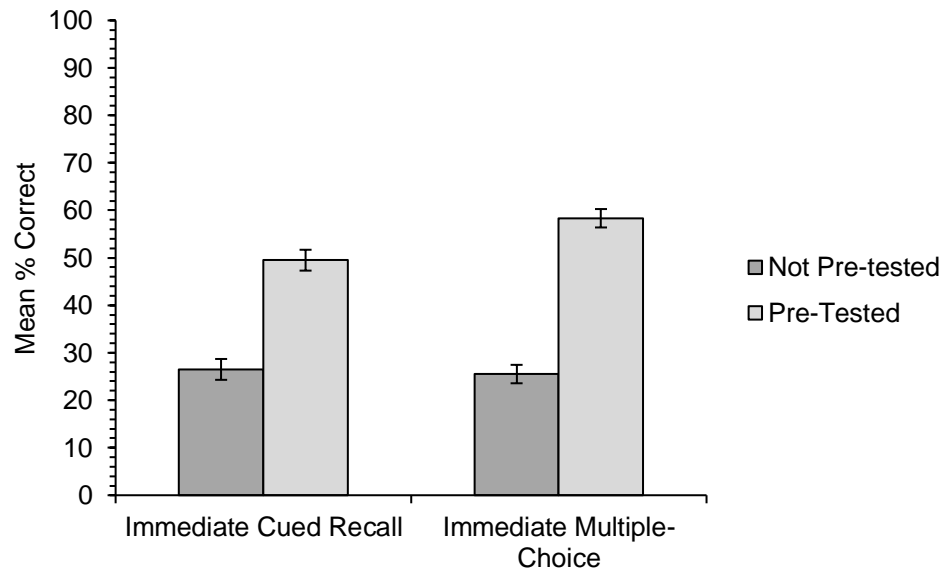
Figure 14. Experiment 4. Mean percentage of correct responses given by participants in the cued recall test[15] (meanings correctly recalled for the appropriate word) measured at the delayed test. Accuracy on the test is shown for participants whose immediate test was also cued recall, and for those whose immediate test was meaning-to-word matching when items were or were not pre-tested. Error bars show standard error of the subject means adjusted for the within-participants factor of whether items were or were not pre-tested (Cousineau, 2005).

The analysis showed a significant main effect of whether items were or were not immediately tested [$\chi^2(1) = 23.73$, $p < .001$], but no significant main effect of immediate test method [$\chi^2(1) = 0.47$, $p = .491$]. The interaction between these two factors was non-significant [$\chi^2(1) = 3.18$, $p = .074$]. The planned simple effects follow-up analysis showed that there was a significant effect of whether items were or were not immediately tested within the immediate cued recall group [$\chi^2(1) = 8.25$, $p = .004$], and also within the multiple-choice meaning-to-word matching group [$\chi^2(1) = 25.10$, $p < .001$]. (The p-values for these simple effects analyses were compared against a Bonferroni-corrected α of .025.)

Multiple choice meaning-to-word matching

The mean percentage accuracy data for the multiple-choice meaning-to-word matching test (measured in the delayed test session) are shown in Figure 15. Performance on this test was

---

[15] The LME analyses were carried out on the raw binary accuracy data, however mean percentage accuracy data are displayed in the graphs for ease of interpretation.

much higher overall than on the prior test of cued recall of the new meanings. Again, accuracy was lower when items had not been tested immediately after training and was similar for the immediate cued recall group (60.5%) and the immediate multiple-choice meaning-to-word matching group (63.5%). Performance again appeared higher in both groups when items had been tested previously, which was again higher for the immediate multiple-choice meaning-to-word matching group (78.6%) than for the immediate cued recall group (67.5%).
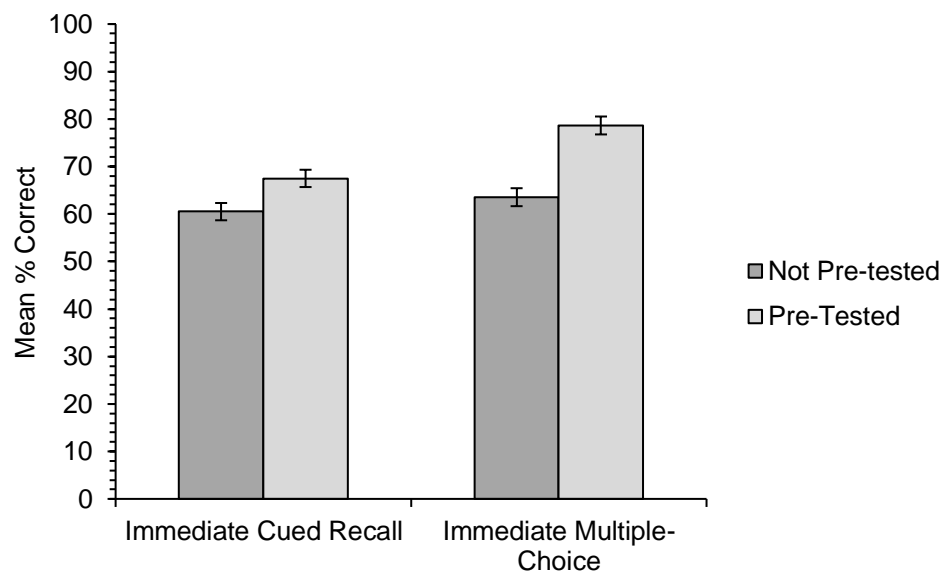
Figure 15. Experiment 4. Mean percentage of correct responses given by participants in the multiple-choice test (meanings correctly matched to the appropriate word) measured at the delayed test. Accuracy on the test is shown for participants whose immediate test was cued recall, and for those whose immediate test was also meaning-to-word matching when items were or were not pre-tested. Error bars show standard error of the subject means adjusted for the within-participants factor of whether items were or were not pre-tested (Cousineau, 2005).

As was also the case for the cued recall measure, the analysis showed a significant main effect of whether items were or were not immediately tested [$\chi^2(1) = 14.54$, $p < .001$] and no significant main effect of immediate test method [$\chi^2(1) = 2.38$, $p = .123$]. The interaction between these two factors was also non-significant [$\chi^2(1) = 2.48$, $p = .116$]. The planned simple effects follow-up analysis showed that the effect of whether items were or were not immediately tested was non-significant within the immediate cued recall group [$\chi^2(1) = 3.31$, $p = .069$]. The effect was, however, significant within the multiple-choice meaning-to-word

matching group [$\chi^2(1) = 10.76$, $p = .001$]. (The p-values for these simple effects analyses were compared against a Bonferroni-corrected α of .025.)

3.4.4 Discussion

Experiment 4 compared the impact of immediate cued recall and recognition tests on the long-term retention of new meanings for familiar words learned incidentally through story reading, in order to tease apart the testing effect observed in Experiment 3. For this experiment participants learned new meanings for familiar words incidentally through reading two stories, and their memory for half of the items they had seen was tested immediately after training. There were two groups, one who had an immediate cued recall test, and one whose immediate test was meaning-to-word matching. All of the participants had their memory for the new word meanings tested 24 hours later using both measures: cued recall followed by meaning-to-word matching. It was predicted that items tested after training would be remember better 24 hours later than those not previously tested, and that the cued recall test would be more helpful for future retention due to the increased retrieval effort required for this test.

Reassuringly, there was no significant main effect of test method group for either delayed measure, showing that the two groups of participants (who had different immediate test methods) performed similarly overall. Indeed, for items not tested immediately after training accuracy was very similar for the two groups on both delayed measures of recall (immediate cued recall: 26.5%; immediate recognition: 25.5%) and recognition (immediate cued recall: 60.5%; immediate recognition: 63.5%). This shows that overall participants' performance was very similar in both groups.

The present study showed that testing memory immediately after training with either a cued recall or meaning-to-word matching test significantly boosted retention as measured at the delayed tests 24 hours later. The sizeable testing effect from Experiment 3 was therefore replicated with each of the two test tasks individually. This finding is consistent with studies that have found a testing effect arising from an immediate cued recall test (Karpicke & Smith, 2012) or an immediate test using multiple-choice questions (Roediger & Marsh, 2005).

The simple effects of testing were examined within each immediate test group separately to determine whether the two different methods of immediate test enhanced performance on the delayed tests individually. The planned simple effects analyses showed that the immediate multiple choice meaning-to-word matching test significantly boosted performance on both of the delayed tests, while the immediate cued recall test only enhanced performance on the delayed cued recall test but not on the delayed meaning-to-word matching

test (the effect was marginally significant at the uncorrected level). The immediate multiple choice meaning-to-word matching test produced a non-significantly larger testing effect on the delayed cued recall test, where previously tested items were recalled 32.8% better than without prior testing, while an immediate cued recall test gave a 23% increase in recall accuracy. Similarly, recognition accuracy at the delayed test was 15.1% higher following an immediate meaning-to-word matching test, but only 7% high with an immediate cued recall test. This is in contrast to the predictions and the retrieval effort hypothesis (Pyc & Rawson, 2009). However, this result is consistent with the findings of Kang et al. (2007) where in one experiment they found that an immediate multiple-choice test was more beneficial than a short answer test when no feedback was given on performance, as was the case in the present study.

While the immediate cued recall test did not significantly improve performance on the delayed meaning-to-word matching test, the immediate meaning-to-word matching test did boost performance on the delayed cued recall test. This crossover benefit of the recognition on later recall of the new meanings for the words discounts the explanation of the benefit of testing as being due entirely to a practice effect of having previously completed the same test task. This demonstrates that knowledge retained from prior testing can be flexibly applied to new contexts of retrieval, in line with the findings of Rohrer et al. (2010) and others.

In summary, Experiment 4 showed that an immediate test of either cued recall or recognition can aid long-term retention of new meanings for familiar words learned incidentally through story reading. The retention benefits of the immediate multiple-choice meaning-to-word matching test were non-significantly larger than for the immediate cued recall test. Furthermore, the immediate multiple-choice meaning-to-word matching test enhanced delayed cued recall of the new word meanings. This suggests that the observed benefit is not simply due to practising the same test task previously. The findings of this experiment and Experiment 3 may explain the good retention seen in the first two experiments of this thesis.

3.5 General discussion

The experiments in this chapter had two aims, the first was to compare the learning of new meanings for familiar words through incidental and intentional learning conditions, and the second was to explore the possible role of testing memory after training in enhancing future long-term retention.

3.5.1 Incidental versus intentional learning

In both Experiment 2 and Experiment 3 participants learned new meanings for familiar words better under intentional learning conditions than incidentally through reading stories. As mentioned in the discussion of Experiment 2, these findings are in line with those of studies that have compared incidental and intentional learning in studies of L2 vocabulary learning (Hulstijn, 1992; Peters et al., 2009) and L1 vocabulary learning with adolescents (Konopak et al., 1987). This effect is likely driven by more attention being directly focussed on encoding the meaning for a word during training through intentional conditions. On the other hand, in the incidental learning conditions participants' attention is also occupied with other aspects of the narrative in the richer context of the stories.

One possible alternative reason as to why the new word meanings were learned better through the intentional learning task is due to the benefits of spaced learning. Spacing stimuli apart has been widely shown to aid learning in comparison to an equal number of study opportunities in which stimuli are more temporally close together (for review see: Dempster, 1996). The definition learning task used in the present study to encourage intentional learning began with reading each definition in turn, followed by two-alternative multiple choice meaning-to-word matching of short, paraphrased versions of the definitions in a pseudorandomised order, followed by a repetition of these two stages of the task to give eight exposures to each item in total (two from reading full definitions, and six from matching the new meanings to the words). The exposures to the new meanings were therefore systematically spaced into the different stages of the task. In the stories, on the other hand, spacing of exposures was not systematic, with new word meanings appearing at naturally-occurring intervals necessary for the stories' narratives. Some of the exposures in the stories were therefore massed (occurring in no more than four consecutive sentences), which is not as conducive to efficient learning (for review see: Dempster, 1996).

Another potential reason for the higher levels of cued recall and recognition accuracy for items learned through the intentional condition is the possibility of an internal testing effect within the training task. The two-alternative multiple choice meaning-to-word matching portion of the definition learning task was similar to the eight-alternative multiple choice meaning-to-word matching task used in the testing phase, which gave rise to a testing effect on its own in Experiment 4. Furthermore, the two-alternative multiple choice meaning-to-word matching task in the intentional learning phase included simple feedback on performance ("correct" or "incorrect"). As mentioned in the introduction to this chapter, feedback enhances the benefit of tests for future retention. Multiple-choice tests have been found to lead to learning of incorrect information from foil responses (Butler et al., 2006; Marsh et al., 2007;

Roediger & Marsh, 2005). However, this risk is greatly reduced with the provision of feedback and a smaller number of foil answers as in the intentional training task used in Experiment 2 and Experiment 3.

As well as the overall differences in performance between the incidental and intentional learning conditions, there were also some differences in long-term retention of items acquired through the different learning modes. After 24 hours, participants in Experiment 2 (and non-significantly in Experiment 3) had forgotten some of the new word meanings learned under intentional conditions, but there was very little forgetting of items learned incidentally through the stories across both of these experiments. This is possibly due to the more semantically rich context of the stories providing participants with additional and more varied cues, which are advantageous for later retrieval of the new meanings. Therefore, while intentional learning conditions were better for more efficient immediate acquisition, incidental learning appears to lead to less forgetting of newly acquired word meanings over time.

3.5.2 The testing effect

A large overall testing effect was found in both Experiment 3 and Experiment 4: retrieval practice following initial exposure boosted retention of new meanings for familiar words. This may therefore explain the high levels of cued recall and recognition accuracy found in Experiment 2 after one day and in Experiment 1 after seven days. This finding adds to the growing literature highlighting the role of testing in aiding vocabulary learning. Future research could explore the relationship between the type of learning materials and the effect of retrieval practice on untested items.

Another aspect of the testing effect for future research is the impact of participants' performance at the immediate test on subsequent retention. As mentioned in the introduction to this chapter, one risk with test-enhanced learning is that as well as boosting retention of previously-presented information, it can also reinforce incorrectly remembered information. In particular, multiple-choice tests may lead to the learning of incorrect information from foil answers (Butler et al., 2006; Marsh et al., 2007; Roediger & Marsh, 2005). On the other hand, other research has shown that under some circumstances testing memory during learning is beneficial for future retention even when initial responses are mostly incorrect (Potts & Shanks, 2014), although in such cases participants often benefit from the provision of corrective feedback. Future research could therefore further investigate this further.

The findings of the present experiments have important methodological implications for studies of word learning. The enhancing effects of retrieval practice on memory are clearly

shown here, and in other previous research. Despite this, some studies considering the impact of other factors such as the importance of sleep for consolidation have somewhat neglected this important aspect. For example, Henderson et al. (2015) compared adult and child participants' explicit memory of new words using cued recall and recognition tests administered both immediately and 24 hours later. They note that for both adults and children "explicit phonological memory was enhanced after off-line consolidation" p.413 (Henderson et al., 2015), although this finding could be attributable to a testing effect. Furthermore, in another study on adults' integration of new words into the mental lexicon over the course of eight days, Tamminen and Gaskell (2013) found that participants whose recall was repeatedly tested throughout the week recalled significantly more novel word meanings at the final test than those who were not previously tested. They reasoned that this this was clear evidence of repeated administration of the recall task preserving participants' explicit memory of the new meanings and protecting them against forgetting (Tamminen & Gaskell, 2013). The testing effect is therefore an important consideration for those studies of the cognitive mechanisms underlying vocabulary learning and retention that include repeated testing of trained words. Studies of sleep and vocabulary learning would benefit from using designs that avoid having multiple test sessions, such as by training different items at different times and testing all items in one final session. Experiment 5 and Experiment 6 in Chapter 4 of this thesis will investigate the possible role of sleep in consolidating new meanings for familiar words without testing participants in multiple sessions in order to avoid contaminating results of potential consolidation with those of a testing effect.

Finally, the findings of Experiment 3 and Experiment 4 have important practical implications for vocabulary learning. Students learning L1 or L2 vocabulary incidentally from reading storybooks or textbooks could benefit from being tested following initial encounters with new word meanings. Testing appears to be effective using either cued recall or multiple-choice methods, so incorporating it as part of a strategy for efficient vocabulary learning could be easy to implement. Tests are often considered solely as tools to assess learning, however they also provide an important opportunity for additional learning and reinforcement of knowledge.

3.5.3 Conclusions

This chapter investigated the comparison between incidental and intentional learning of new meanings for familiar words, as well as the role of tests immediately after initial learning in maintaining memory of new word meanings over time. In line with previous findings in the literature on L2 vocabulary learning and children's L1 vocabulary learning, the first two

experiments in this chapter demonstrated that new word meanings are learned more efficiently under intentional learning conditions than incidentally through story reading. However, there was also some evidence of less forgetting of items learned through stories, suggesting that word meanings learned in a more semantically rich context could be retained better. The second two experiments in this chapter showed that testing memory aids future long-term retention of new meanings for familiar words acquired under either incidental or intentional conditions. Both cued recall and recognition tests enhanced retention, but multiple-choice tests gave better performance in the present context in which no feedback was given on performance. Furthermore the testing effect transferred across test tasks: immediate meaning-to-word matching improved accuracy on the delayed cued recall test, so the effect is not restricted to benefitting the previously completed test task. Testing memory following initial exposure is therefore a powerful way to improve learning and long-term retention of vocabulary knowledge.

Chapter 4: Overnight consolidation of new meanings for familiar words

4.1 Introduction

The previous chapter investigated the impact of learning conditions and prior testing on the acquisition and long-term retention of new meanings for familiar words. The testing effect is thought to be beneficial for memory due to participants' online retrieval of related knowledge aiding the formation of adaptable hippocampal-neocortical representations (Antony et al., 2017). A slower route by which these representations may be formed and strengthened is through offline consolidation during sleep. As adults continually learn new meanings for familiar words, they must integrate information about the newly-learnt meanings with their existing knowledge about the prior meanings of words. With such an abundance of information stored in the mental lexicon, it is a challenge for the adult learner to manage it all, and acquire new meanings for words whilst preserving their knowledge of the pre-existing meanings. This chapter explores how this may be achieved through two experiments that investigate the potential role of overnight consolidation during sleep in adults' learning of new meanings for familiar words.

4.1.1 Complementary Learning Systems

The Complementary Learning Systems (CLS) theory of word learning (Davis & Gaskell, 2009) provides an explanation for how new vocabulary may be integrated with pre-existing knowledge. Davis and Gaskell (2009) combined the CLS model of learning and memory (McClelland, McNaughton, & O'Reilly, 1995) with behavioural and neural findings from studies on spoken word form learning, to develop an account of how adults process and learn new words. They describe how words are initially encoded into episodic memory in the hippocampus, and become quickly learnt and familiar, but only after a period of long offline consolidation during sleep does knowledge of these new words become integrated into semantic memory in the neocortex. It is only once words have become assimilated into the mental lexicon that they are more rapidly recognised and able to compete with existing similar word forms during word recognition, for example in a lexical decision task (Tamminen & Gaskell, 2013; Leach & Samuel, 2007).

It is well established that recognition of a spoken word entails lexical competition between phonologically neighbouring words that become activated by overlapping phonological input, such as with *captive* and *captain* (Gaskell & Dumay, 2003; Luce & Pisoni, 1998; Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994). Similarly, in the semantic ambiguity literature, Rodd et al. (2002) have described how recognition of a homonym involves semantic competition between the word's semantically unrelated

meanings (e.g., *bark*—tree/dog). Lexical or semantic competition are examples of lexical engagement (Leach & Samuel, 2007), as they demonstrate how the activation of one word can have an impact on the activation of another word. Consolidation of a word or meaning is a necessary precursor to competition. Previous research (e.g., Gaskell & Dumay, 2003) has shown a dissociation between lexical configuration (that is familiarity with a word's sound, spelling, meaning, or grammatical usage; Leach & Samuel, 2007), which begins to develop immediately, and more slowly-emerging lexical engagement, at which point a word interacts with other items in the mental lexicon. Consolidation of new words or meanings can therefore be assessed using implicit measures of memory that probe lexical (e.g., Gaskell & Dumay, 2003) or semantic (e.g., Maciejewski, Rodd, Mon-Williams, & Klepousniotou, 2018) competition between novel words or meanings and their pre-existing lexical competitors. On the other hand, explicit measures of memory (e.g., recognition and recall) assess overall knowledge, or lexical configuration, of words but are not able to ascertain whether that information has been consolidated into semantic memory (Henderson et al., 2015). It is therefore important for studies of word learning to include both explicit and implicit measures of memory for new words and/or meanings in order to fully understand the process by which new entries are added to the mental lexicon.

The CLS model proposes an active role for sleep in the consolidation of novel vocabulary items into semantic memory. However an alternative account is that sleep offers a passive benefit to memory in that it protects against interference due to a lack of encoding of new information (Jenkins & Dallenbach, 1924). There is a longstanding debate surrounding the nature of the benefit of sleep for memory (for a review see: Ellenbogen, Payne, & Stickgold, 2006), however recent research has provided strong evidence for an active role for sleep. For example manipulating brain oscillations has been shown to enhance memory of word pairs (Ngo, Martinetz, Born, & Mölle, 2013), and targeted memory reactivation has been found to facilitate consolidation of picture-location associations (Cairney, Durrant, Hulleman, & Lewis, 2014). However a passive account of the benefit of sleep should not be discounted. An active role of sleep in consolidation of newly-acquired words and/or meanings may only be verified through implicit measures that probe competition as a result of integration with prior knowledge.

4.1.2 Consolidation of word forms

The CLS theory of how adults learn new words has been largely based around evidence from studies of spoken word form learning. For example, in one of the earlier studies in this area (Gaskell & Dumay, 2003), participants learned novel words that were artificial phonological

neighbours to existing words (e.g., *cathedruke* for *cathedral*). Participants' knowledge of these new words was then tested using a recognition test of the new word form, and lexical decision or pause detection tests of the existing words as measures of competition with the new words due to their lexicalisation (Takashima, Bakker, van Hell, Janzen, & McQueen, 2014), immediately and eight days later. Gaskell and Dumay (2003) found inhibited access for the existing words (due to competition for access from the new word forms) at the delayed test but not immediately after training, suggesting that offline consolidation is required for words to become integrated into the mental lexicon (Tamminen & Gaskell, 2013). Further studies of this nature have replicated these findings (e.g., Davis, Maria, Betta, Macdonald, & Gaskell, 2009; Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010). Several of these studies have also specifically shown the importance of sleep for the consolidation process, by dissociating sleep from the simple passage of time (Dumay & Gaskell, 2007), and showing associations between specific components of sleep, such as sleep spindles, and lexical integration (Tamminen et al., 2010). This body of work therefore provides evidence supporting the CLS account of word learning, and suggests that sleep may be a special state for lexical consolidation.

However, several recent studies have provided evidence for the lexicalisation of new words without sleep. For example, Kapnoula, Packard, Gupta, and McMurray (2015) trained participants on short non-words that differed from real words on the final phoneme (e.g., *jod* and *job*), and tested for inhibition effects from these novel word forms on the existing words using the visual world eye-tracking paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In the experiment they used a phoneme splicing manipulation to amplify competition between the new words and their pre-existing competitors. This competition was measured by examining participants' eye movements towards a picture representing the existing word which was presenting alongside pictures denoting three unrelated filler words (one with some phonological overlap). They found that the new words were able to compete with the existing words for access immediately after training, without any consolidation (Kapnoula et al., 2015). Kapnoula and McMurray (2015) later provided further evidence that this competition seems to derive from lexicalised representations of the novel words, rather than from episodic memories, in contrast to CLS predictions. In a replication of the previously mentioned *cathedruke* study, Lindsay and Gaskell (2013) also found that words could be integrated immediately, without a period of sleep-based consolidation, when the novel words were repetitively trained using spaced learning alongside exposure to their existing competitor words. Therefore, under certain circumstances immediate integration of novel word forms seems to be possible without consolidation during sleep.

4.1.3 Consolidation of word meanings

However, studies of learning word forms in isolation are not ecologically valid as word forms are not learned without meanings in everyday life, and this may lead participants to engage in more deliberate memorisation that would occur in natural lexical learning. Importantly, increasing the richness of information about a novel word by adding semantic information has also been shown to differentially engage the complementary memory systems during lexicalisation (Takashima et al., 2014), compared to word form learning alone. It is therefore necessary to consider the implications of the CLS account for the learning of novel word meanings, as it is unclear whether the same mechanisms that are involved in learning word forms are also involved in the acquisition of new word meanings. Existing research has combined learning of new word forms with corresponding semantic information, for example by training participants on pronounceable non-words with invented picturable meanings (Clay, Bowers, Davis, & Hanley, 2007), meaningful affixes attached to existing words (Tamminen, Davis, Merkx, & Rastle, 2012), or low-frequency existing words (Van Der Ven, Takashima, Segers, & Verhoeven, 2015). Each of these three studies found that effects of semantic integration only arose following a period of overnight consolidation, which suggests that information about a word's meaning also requires time to become integrated into semantic memory (Van der Ven et al., 2015), and is consistent with the CLS theory.

Nevertheless, as with word form learning alone, evidence has also been found for the semantic integration of novel words and their meanings without sleep-dependent consolidation, such as in the area of language production (Oppenheim, 2015). Others have suggested that the method of encoding of novel words and their meanings can greatly impact upon subsequent semantic integration (Coutanche & Thompson-Schill, 2014). Coutanche and Thompson-Schill (2014) showed that using the fast-mapping learning procedure (whereby participants are forced to infer meaning by process of elimination) enabled immediate lexical integration, while the more traditional explicit encoding procedure produced integration only after consolidation. Offline consolidation may therefore not be a prerequisite for the integration of new knowledge when the learning conditions encourage connections to be formed online between the new information and existing knowledge (Fang et al., 2016), as is the case with fast mapping (Coutanche & Thompson-Schill, 2014), spaced learning (Lindsay & Gaskell, 2009), and test-enhanced learning (Antony et al., 2017).

However, all of the aforementioned studies of word meaning learning and consolidation combined the acquisition of a new meaning with simultaneous acquisition of a novel word form, which is different to learning a new meaning for an existing word form that already has semantic information attached to it. This is an important distinction, because if both are

learned together it is hard to disentangle whether the consolidation effects reflect learning of the form or the meaning, or both. Furthermore, Rodd et al. (2012) showed that properties of the existing meaning of a word (its semantic relatedness to a novel meaning) can affect the ability to learn a new meaning for that same word. Learning new related and unrelated meanings likely involves different learning mechanisms, as learning new related information may promote the online reactivation of related knowledge, bypassing the need for offline consolidation (Antony et al., 2017; Fang et al., 2016), while this would be of little use for learning new unrelated meanings for familiar words. So far only a few studies (Fang & Perfetti, 2017; Fang et al., 2016; Maciejewski et al., 2018; Rodd et al., 2012) have looked at the effects of making an unambiguous word into an ambiguous one by assigning it a novel invented meaning. These studies did not examine any potential effects of overnight consolidation of the new meanings directly; while Fang and Perfetti (2017) did measure explicit memory immediately after training and one week later, they only measured meaning integration (using an ERP measure) at the immediate test and not following sleep.

4.1.4 Measures of meaning integration

Various different implicit behavioural measures have been used to assess integration of new word meanings into semantic memory. For example, Rodd et al. (2012) assessed the impact of newly-learned semantically related and semantically unrelated meanings for familiar words on recognition of the words using a lexical decision task. Their premise was that if no integration of the new meanings had taken place, then participants would rely only on their prior knowledge of the words when responding in the lexical decision task. As such there would be no difference in responses between words whose new meanings were semantically related to the pre-existing meaning and those whose new meanings were semantically unrelated. They found that words with semantically related new meanings were recognised faster than words with new unrelated meanings, therefore indicating that the new word meanings had been integrated with prior semantic knowledge of the words. However, this relatedness effect only emerged when a more intensive training regime was used (Experiment 3; Rodd et al., 2012), as otherwise they had not been sufficiently integrated into participants' lexicons to cause interference in online word recognition (Rodd et al., 2012). Similarly, Van der Ven et al. (2015) used a primed lexical decision task to assess the semantic integration of newly-learned words and their meanings. They tested whether novel words primed recognition of semantically related words before and after participants had learned new meanings for them (Van der Ven et al., 2015). The results showed that the priming effect was not present before word learning, but it emerged after a 24-hour delay (Van der Ven et al., 2015). However, while

these two studies demonstrate that lexical decision tasks do involve semantic processing to some extent, recognising a word may not necessarily require access to its meaning. Indeed, previous studies (e.g., Azuma & Van Orden, 1997) have revealed some uncertainty in the degree of semantic access involved in making lexical decisions. For the present research it is particularly important that a task selected to measure integration of new word meanings entails semantic access, because competition between unrelated meanings of homonyms occurs within the semantic level of representation (Rodd et al., 2004).

Other behavioural tasks have been used to assess consolidation that specifically require access to a word's meaning. (Gaskell, Cairney, & Rodd, 2018) used a word association task to investigate whether consolidation during sleep affects the processing of ambiguous words. Participants were primed on the subordinate meaning of ambiguous words (e.g., "pen" – animal enclosure; Gaskell et al., 2018) through listening to sentences. After a period of either two or 12 hours' sleep or wake (Experiment 1), or 12 hours' sleep and 12 hours awake (Experiment 2), participants' meaning preference was assessed using a word association task in which they were required to type the first word that came to mind in association with a word (e.g., "pen"; Gaskell et al., 2018). Gaskell et al. (2018) found that participants were more likely to choose the primed subordinate meaning after a period of sleep than a period of wake, and sleep seemed to protect the primed meaning from external interference when it closely followed priming. However, word association can only serve as a measure of consolidation when both meanings of a word are already familiar. Furthermore, it is a relatively slow offline measure (Cai et al., 2017) that cannot measure processing of the different meanings of an ambiguous word individually (Betts, 2018). Word association is therefore of little use for assessing consolidation of new meanings for familiar words.

On the other hand, speeded semantic relatedness judgement has previously been used to investigate the impact of newly-learned meanings for previously unambiguous words on processing of the pre-existing meaning of the words (Maciejewski et al., 2018). In two experiments Maciejewski et al. (2018) trained participants on new, fictitious meanings for words in an intensive training regime spread across four days, similar to the procedure used by Rodd et al. (2012). The participants completed a speeded semantic relatedness decision task both before and after learning the new meanings for the words, and their cued recall of the new meanings was assessed. In the semantic relatedness decision task, participants were presented with the trained words as target words and had to decide whether a subsequent probe word was related or unrelated to that word, with probes designed to test processing of the existing meaning and not the new meaning of the words (Maciejewski et al., 2018). For example *sip* was given the semantically related new meaning of "a small amount of computer data", and the probe used for this word, *juice*, was only related to the pre-existing meaning but

not the new meaning assigned to the word. They found that after training participants' responses were slower for the words they had learned new meanings for, but not for untrained control words, and the effect was larger for words whose new meaning was semantically unrelated to the existing meaning (Maciejewski et al., 2018). This demonstrates competition arising between the newly-learned and well-established meanings of words as a result of integration of the new meanings into the lexicon over the four days of training (Maciejewski et al., 2018). This was the first research to use semantic relatedness judgement as a measure of consolidation of new meanings for familiar words (Maciejewski et al., 2018). While Fang et al. (2016) also previously used a semantic relatedness judgement task in a study of learning new meanings for known words, they probed the new meanings to assess whether existing meanings affected retrieval of the new meanings. Such a task cannot evidence lexicalisation of new meanings for words as it does not necessarily entail accessing semantic memory; participants could rely solely on episodic memory of learning new meanings in the experiment in order to successfully complete the task (Fang et al., 2016). For the semantic relatedness judgement task to evaluate consolidation of new meanings for words it must therefore probe processing of the pre-existing meaning to see whether it is modulated by acquiring knowledge of a new meaning for the same word.

4.1.5 Chapter overview

The experiments in this chapter investigate the impact of overnight consolidation during sleep on learning new meanings for familiar word forms acquired incidentally through story reading. Experiment 5 compares participants' memory of new meanings for familiar words after 12 hours including sleep to 12 hours of wake. Building on this, Experiment 6 compares participants' memory of new word meanings trained either 24 hours or 12 hours prior to test, with participants tested either in the morning or in the evening to try to tease apart active and passive benefits of overnight sleep. In both experiments explicit knowledge of new meanings is assessed using tests of cued recall and recognition (multiple-choice meaning-to-word matching). An implicit measure of reaction time on a semantic relatedness judgement task is used to assess semantic integration of the new meanings by examining competition effects arising between the new and existing meanings of the words.

4.2 Experiment 5: Overnight consolidation after a single study session

4.2.1 Introduction

The aim of Experiment 5 was to examine whether overnight consolidation is beneficial for the learning of new meanings for familiar words, as has previously been shown for learning new word forms (e.g., Tamminen & Gaskell, 2013). The experiment used a between-groups design in which participants were trained on new meanings for familiar words through reading stories either in the evening or the morning (see Figure 16). This was then followed by a delay of 12 hours of either sleep or wake, and then by a test session. In the test session consolidation of the new meanings was assessed using an implicit measure of reaction times on a speeded semantic relatedness judgement task, which probed semantic competition between the new and pre-existing meanings of the words. Participants' explicit knowledge of the new meanings for the words was tested through cued recall and a multiple choice meaning-to-word matching recognition test. It was predicted that participants who had slept would show inhibition (slower reaction times) for the trained words in the implicit reaction-time measure, due to competition arising between the new and old meanings, which would not be the case for those who had not slept. The predictions of the CLS model for the effect of sleep on explicit memory are less clear. Explicit memory measures assess overall knowledge of words (lexical configuration) which has been shown to dissociate from lexical engagement (Gaskell & Dumay, 2003; Leach & Samuel, 2007), however several studies have found enhanced explicit memory for word forms following overnight sleep (Dumay, Gaskell, & Feng, 2005; Henderson et al., 2015; Takashima et al., 2014; Tamminen et al., 2010). It was predicted that participants who had slept would have better explicit memory of the new word meanings than those who had not slept.
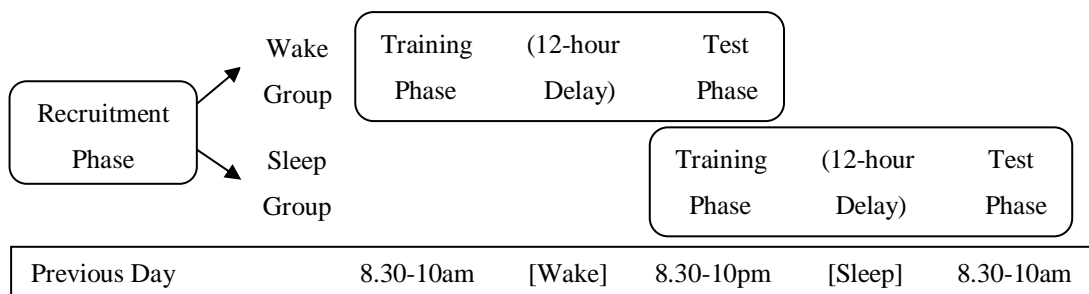
Figure 16. Diagram demonstrating the procedural design for the two groups in Experiment 5.

4.2.2 Method

Participants

The participants included in the study were eighty-four adults (age: $M = 31.4$ years, $SD = 8.6$, range = 18-49; 39 females). All participants were monolingual native speakers of British English who had not been diagnosed with any reading or language disorders, and who had not taken part in any of the previous experiments. None of the participants reported having been diagnosed with a sleep disorder or were taking any medication that could affect their sleep. Participants were recruited through the website Prolific Academic (Damer & Bradley, 2014) and were paid for their participation upon completion of the whole experiment (£8 in total).

In addition to the 84 participants included in the study, 56 participants started but failed to complete all sessions of the experiment so no data was analysed for them. Thirteen further participants were excluded for reporting having a sleep disorder or currently taking medication that could disrupt their sleep. A further 15 participants were excluded due to getting more than two of the multiple-choice comprehension questions wrong on the stories. Upon inspection of the data from the training session, four participants were excluded due to being outliers in their mean reading speeds (faster than 624.0 words per minute, two standard deviations above the mean). Eleven participants were excluded due to misunderstanding the instructions of the cued recall test; and one participant was excluded due to having low accuracy on the semantic relatedness judgement test task (less than 87.8%, three standard deviations below the mean). Excluded participants were replaced to obtain the total of 84 participants included in the study.

Materials

Novel word meanings and short stories

The stimuli for the experiment were the same 16 words with novel semantically unrelated meanings as used in the previous experiments (see Appendix A for a list of the stimuli), which had been incorporated into the four separate short stories (see Appendix B). The same longer paraphrased versions of the definitions as used in the previous experiments were again used in the multiple-choice meaning-to-word matching test in this experiment (see Appendix D).

Stimuli for semantic relatedness judgement test task

Stimuli for the semantic relatedness judgement task comprised the 16 stimulus words mentioned previously that had been paired with novel meanings. Each of these stimulus words

was paired with both a semantically related probe word (e.g., *hive-honey*) and a semantically unrelated probe word (e.g., *hive-bicycle*) for use in the semantic relatedness judgement task (see Appendix E for the full list of target words with their semantically related and semantically unrelated probes). The majority of the semantically related probes ($n = 10$) and unrelated probes ($n = 12$) were selected from those used by (Maciejewski et al., 2018) paired with the same target words in a semantic relatedness judgement task. The remaining semantically related probes ($n = 6$) were selected from the Edinburgh Association Thesaurus (EAT; Kiss, Armstrong, Milroy, & Piper, 1973), and the remaining unrelated probes ($n = 4$) were selected from the remaining set of probe words from (Maciejewski et al., 2018) that had been used with other target words. The degree of semantic relatedness between each target word and its corresponding related and unrelated probes was determined using Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998). The mean LSA measure by pairwise comparison between the semantically related probes and targets was 0.4 ($SD = 0.2$), and between the semantically unrelated probes and targets it was 0.06 ($SD = 0.1$). All of the probe words were nouns, and the related and unrelated probes were matched as closely as possible to both the target words and to each other in terms of their frequency and word length, and were also similar in terms of their number of senses and number of semantic associates (see Table 2). Care was also taken to ensure that the probe words were not semantically related to the novel meanings of the words, and none of the probe words appeared in any of the stories.

	Related Probes	Unrelated Probes
Example	*dawn-dusk*	*dawn-basket*
No. of Letters	5.19 (1.17)	5.44 (1.15)
Frequency (per mil.)	26.28 (19.38)	17.59 (12.05)
Frequency (log-transf.)	3.58 (0.41)	3.45 (0.32)
WordNet Senses	5.56 (3.44)	5.50 (4.66)
Wordsmyth Senses	6.19 (4.05)	4.88 (2.75)
No. Semantic Associates	14.69 (5.26)	14.27 (3.69)[16]
Target-Probe Relatedness	0.44 (0.18)	0.06 (0.07)

Table 2. Descriptive statistics for the lexical and semantic properties of the probe words used in the semantic relatedness judgement task in Experiment 5. The means for each measure are displayed in the table, with standard deviations given in parentheses. The words frequency data reported are the SUBTLEX-UK word frequencies in occurrences per million and log-transformations of the raw word

---

[16] There was no data for one item (*alien*, which was the semantically unrelated probe for *cake*).

frequencies (log10[raw frequency+1]) (Van Heuven et al., 2014). Word sense data were taken from the WordNet (Fellbaum, 1998) and Wordsmyth (Parks et al., 1998) dictionaries. The number of semantic associates counts come from Nelson et al. (2004). The target-probe semantic relatedness values are Latent Semantic Analysis (LSA) estimates (Landauer et al., 1998).

Additionally, 16 fillers were selected from the control words used by Rodd et al. (2012) and Maciejewski et al. (2018) that were also matched to the stimuli in terms of their frequency and word length. Half of these fillers (selected at random) were paired with two semantically related probes, and the other half were paired with two semantically unrelated probes. This was done in order to prevent a predictable pattern in the task whereby each target word seen would appear once with a related probe followed by an unrelated probe, or vice versa, and the inclusion of these fillers was intended to prevent participants from being able to anticipate the correct response before seeing the probe word on a given trial. The semantically related and unrelated probes for the fillers were selected in the same way as before.

An extra eight unmatched fillers were selected to serve as a practice block before the start of the main experimental task. These fillers were paired with probes to give the same distribution as the trials in the main experiment: half were paired with both a related and unrelated probe, and half were paired with either two related or two unrelated probes. Another eight extra unmatched fillers, with the same distribution of target-probe pairings as the practice block, were selected to appear at the beginning of the experimental task blocks in order to accustom participants to the speed and rhythm of the task.

The trials for the experimental task were split into two separate blocks of 40 trials each (eight 'starting fillers', followed by 16 experimental trials intermixed with 16 matched filler trials), with a brief break between the two blocks. Each target word (stimulus or filler) appeared once in each of the two blocks, paired with one of its two probe words in the first block, and the other in the second block. For the stimulus items, half of the targets (selected at random) appeared with their semantically related probe in block 1 and then with their semantically unrelated probe in block 2; the other half appeared with their semantically unrelated probe in block 1, then with their semantically related probe in block 2. The order of the two blocks was counterbalanced across participants. The order of trials within each of the two experimental blocks (and the starting fillers at the beginning of each of the two blocks) was randomised for each participant.

Design

To ensure that each new word meaning appeared roughly an equal number of times in each condition, and that the order of the two blocks in the semantic relatedness judgement task was counterbalanced across participants, eight versions of the experiment were set up. Half of the participants ($N = 42$) were randomly assigned to the wake group, and the other half ($N = 42$) to the sleep group, which determined at what times they would be required to complete the training and test sessions of the experiment. Almost exactly half the participants ($N = 41$) were trained on the set of words occurring in stories 1 and 4, and the remaining participants ($N = 43$) were trained on the set of words occurring in stories 2 and 3; each participant was therefore trained on half the total number of stimuli (eight items). Finally, the order of the two blocks of trials in the semantic relatedness judgement task was counterbalanced across participants in order to minimise any potential order or repetition effects of seeing each stimulus twice (once with a semantically related probe, and once with an semantically unrelated probe). Participants were randomly assigned to one of these eight versions in the recruitment phase of the experiment.

Procedure

The experiment was carried out by participants online using Qualtrics Survey Software (Qualtrics, 2015) , with the Qualtrics Reaction Time Engine (QRTE; Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015). Figure 16 shows a schematic of the experiment, with the timescale for the training and testing of the two experimental groups.

Recruitment phase

In the first phase of the experiment, participants were asked to provide some demographics details and to commit to taking part in all three sessions of the experiment. Participants were then randomly assigned to one of the eight versions of the experiment, which determined whether they were to be part of the wake group or the sleep group. They were then given the times for their two subsequent sessions starting the following day at either 8.30-10am and 8.30-10pm for the wake group, or 8.30-10pm and 8.30-10am the following morning for the sleep group. Participants were not told that the purpose of the experiment was to learn new word meanings, and were not aware that their memory would be tested. Instead they were told the cover story that the purpose of the experiment was to investigate reading ability and comprehension of texts at different times of day.

Training phase

During the training phase, participants each read the two short stories that they had been allocated according to the version of the experiment that they had been assigned to (either Stories 1 and 4, or 2 and 3). The procedure for reading the stories and answering the comprehension questions was the same as described in the previous experiments in this thesis (see Chapter 2 for details). After completing the first story, participants were given a brief break of 20 seconds before they were allowed to continue on to begin reading the second story. As with all of the previous experiments, the purpose of the comprehension questions was to check that the participants had read the stories carefully and fully processed the meaning, and therefore served as exclusion criteria. Participants were excluded if they got more than one of the five comprehension questions wrong on either of the stories they read: 15 participants were excluded for this reason.

Testing phase

Semantic relatedness judgement task

The first task of the test phase was the semantic relatedness judgement task. Participants were presented with the stimuli and filler items one at a time in the centre of the screen. Each individual trial began with a fixation cross presented for 500ms, followed by the target word for 500ms, then a fixation cross presented for another 500ms, and finally the probe word was presented until a response was given. The participants' task was to decide whether the target and probe word were semantically related (e.g., *hive-honey*) or not (e.g., *hive-bicycle*). Participants were not told which meaning of the target word they should attend to, and were instructed to try to respond as quickly and accurately as possible. They indicated their choice with a "yes" response (by pressing the "j" key), or a "no" response (by pressing the "f" key). If a response was not given until after the probe had been onscreen for 2000ms, then a message was displayed to tell the participant that their response was too slow and that they should respond more quickly.

Before beginning the experimental task, participants first completed a practice block of 16 trials. The purpose of the practice block was to familiarise participants with the task and to provide feedback on their speed and accuracy. Following each practice trial, a feedback screen informed them whether their response had been correct or incorrect. Additionally, if a response was slower than 2000ms, another feedback message told participants that their response was too slow and they should respond more quickly. Following the practice block, participants proceeded to the first of the two experimental blocks, which each began with eight starting

fillers in order to accustom participants to the speed and rhythm of the task. Following this there were 16 experimental trials and 16 matched filler trials in each of the two blocks, the order of which was randomised separately for each participant. There was a break of at least ten seconds between the two experimental blocks.

Cued recall test

Immediately following the semantic relatedness judgement task, participants were given a cued recall test for all of the new meanings for the words that they had been trained on in the previous session. The procedure for this test was the same as for the previous experiments. Participants were presented one at a time with the eight stimulus words for which they had been trained on a novel meaning; they were instructed to recall the appropriate new meaning for each of the words that they had encountered in the stories and type it into a blank text box. They were asked to give as much detail as they could and to try to answer in full sentences even if they were unsure of their answer. If they could not remember anything about the new meaning for the words, then they were instructed to type "don't know". The order of presentation of the word cues was randomised for each participant, and participants were only tested on the eight items that they had been trained on.

Perhaps due in part to the cued recall task directly following the semantic relatedness judgement task without a break, and due to not having read the instructions for the second task carefully, a number of participants did not do what was asked of them in the cued recall task. Eleven participants gave for all of their answers either one of the probe words that was paired with the stimulus word in the preceding task, or the real pre-existing meaning of the word. As mentioned previously, these 11 participants were therefore excluded on the basis of misunderstanding the instructions of the cued recall test.

Multiple choice meaning-to-word matching test

In the final test task, as in previous experiments, participants were presented one at a time with short sentences giving definitions of the novel word meanings that they had been trained on through the stories. The sentences omitted the words to which they were referring, and for each novel meaning participants were asked to select the word that they thought matched the definition from a list of all eight of the stimulus words for which they had been trained on a new meaning. The order of the eight words to choose from was randomised for each test item, as was the order of presentation of the new meaning definitions. As with the cued recall test, participants were only tested on the eight items that they had seen during the training phase.

4.2.3 Results

Stanford Sleepiness Scale

The results for the Stanford Sleepiness Scale (SSS; Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973) measured at the test session were analysed using a Wilcoxon-Mann-Whitney test. There was no significant difference in SSS score between the wake group (*Mdn* = 3) for whom the test session was in the evening, and the sleep group (*Mdn* = 3) for whom the test session was in the morning [W = 798, *p* = .437].

Analysis procedure

Responses for the cued recall test were coded by the experimenter blind to condition. The procedure for coding the responses was the same as for previous experiments, using simple binary accuracy coding ("1" for correct and "0" for incorrect). The data from the multiple choice meaning-to-word matching task were coded in the same way ready for the analysis.

Accuracy was very high overall for all stimulus items in the semantic relatedness judgement task. For one item ("foam") accuracy was slightly lower than three standard deviations below the grand mean for all items (89.88%), however as it was only 0.29% lower than this the item was kept in. The raw reaction time (RT) data from the semantic relatedness judgement task were pre-processed prior to analysis. RTs for incorrect trials were removed from the data (2.6% of all trials), and RTs were trimmed out of the data if they were faster than 300ms or slower than 2500ms (0.2% of the remaining trials). The main analysis was only of the correct related trials, for which the participants had correctly responded that the target and probe words were semantically related.

The data from all three test tasks were analysed using linear mixed effects (LME) models (using the lme4 package (version 1.1-7; Bates et al., 2015) and R statistical software (version 3.0.2; R Core Team, 2017). One model was created to analyse the data from the cued recall test, the multiple choice meaning-to-word matching test, and the semantic relatedness judgement task separately.

The binary accuracy data from both the cued recall test and multiple choice meaning-to-word matching test were analysed using logistic LME models. These two models contained random effects by participants and items (with a random intercept and slope for group by items, and a random intercept by participants), and a fixed effect for group (two levels: sleep group or wake group). The contrasts were defined using deviation coding (sleep group: 0.5 vs. wake group: -0.5).

The reaction time data from the semantic relatedness judgement task were analysed using an LME model containing random effects by participants and items (with a random intercept and slopes for group, training condition, block position, the three 2-way interactions, and the 3-way interaction by items; and a random intercept and slopes for training condition, block position, and the interaction by participants). The model also contained fixed effects for group (two levels: sleep group or wake group), training condition (2 levels: trained items or untrained items), and block position in the task (two levels: first or second). The contrasts were defined using deviation coding for group (sleep group: 0.5 vs. wake group: -0.5), training condition (trained items: 0.5 vs. untrained items: -0.5), and block position (first block: -0.5 vs. second block: 0.5). The contrasts for the interactions were created by multiplying the contrasts for the appropriate variables together.

The first attempted fit for each of the models used the maximal random effects structure, as per guidelines outlined by Barr et al. (2013). For the two logistic LME models, the "bobyqa" optimiser was used for consistency with the analyses for the previous experiments in this thesis. The models for the cued recall and multiple choice meaning-to-word matching measure converged with maximal random effects, and so these were the final models used for the analyses of these measures. For the analysis of the RT data from the semantic relatedness judgement task the maximal model also converged, and so was used as the final model for the analysis. However, the assumptions of homoscedasticity and normality were violated in the raw reaction-time data, so the data were $\log_{10}$- and inverse-transformed (invRT = 1000/rawRT) and compared with the raw RT data. Histograms showing the distributions of these data and scatterplots of the residuals vs. fitted values were created to compare the raw, log-, and inverse-transformed data (see Appendices G and H). The inverse-transformed RTs met the assumptions of homoscedasticity and normality most closely and were therefore used for the analysis. Significance of the fixed effects and interactions were assessed using likelihood ratio tests which compared the full models to models with only the factor or interaction of interest removed (leaving in any other interaction involving that factor or interaction, and leaving the random effects structure intact).

Cued recall of novel meanings

The mean percentage accuracy data for the cued recall test (Figure 17) show that overall accuracy was low, at less than 50% in both groups. However, the sleep group correctly recalled more of the novel word meanings (47.9%) than the wake group (36.3%). There was a significant main effect of group [$\chi^2(1) = 4.13$, $p = .042$], with the sleep group correctly recalling more of the novel word meanings than the wake group.
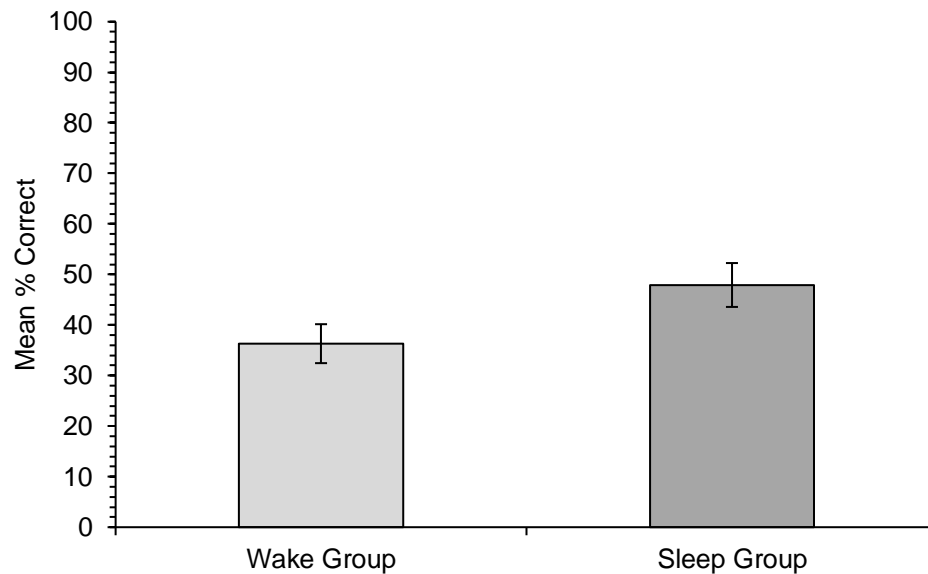
Figure 17. Experiment 5. Mean percentage of correct responses by subjects on the cued recall test[17] (meanings correctly recalled for the appropriate word) for participants in each of the two groups. Error bars show standard error for subject means.

Multiple choice meaning-to-word matching

The mean percentage accuracy data for the multiple choice meaning-to-word matching test (Figure 18) show that overall accuracy was higher than in the cued recall test, but was not near ceiling. The pattern of the data was the same as for the cued recall test: mean accuracy on the task was higher for the sleep group (74.1%) than for the wake group (60.1%). There was a significant main effect of group [$\chi^2(1) = 7.01$, $p = .008$], with the sleep group correctly matching more of the novel meanings with the appropriate words than the wake group.

---

[17] NB. The LME analyses were not carried out on the percentage data, however percentage data are displayed in the graphs for the cued recall test and multiple choice test results for ease of interpretation.
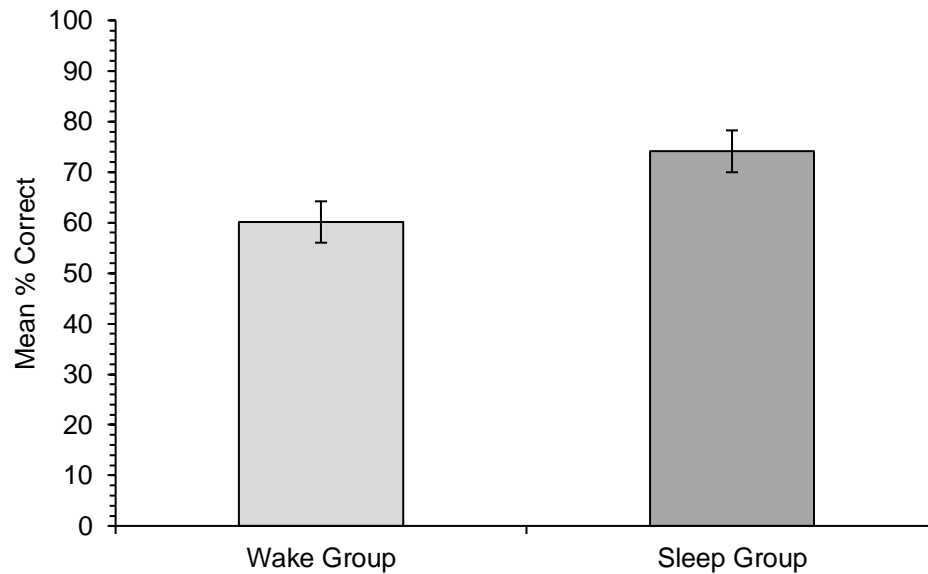
Figure 18. Experiment 5. Mean percentage of correct responses by subjects on the multiple choice meaning-to-word matching test (words correctly matched with the appropriate meaning) for participants in each of the two groups. Error bars show standard error for subject means.

Semantic relatedness judgement

Participants' accuracy in identifying target words and probes as semantically related was very high overall ($M = 95.6\%$, $SD = 5.1\%$) and only differed by a maximum of 1.8% between the different conditions[18]; the accuracy data was therefore not analysed further. The reaction data from the semantic relatedness judgement task (Figure 19) showed that the sleep group appeared to respond faster overall compared with the no sleep group. There appeared to be a trend in the data for participants in both groups responding slightly slower to trained items than untrained items on average. However, this mean difference was only small and the error bars (showing standard error for subject means) were highly overlapped. Furthermore, the pattern of the mean reaction time for trained and untrained items appeared to be the same for both groups, and there was no indication of an interaction between the two variables as had been predicted.

---

[18] Mean percentage accuracy on the semantic relatedness judgement task for the wake group was 96.4% ($SD = 18.36\%$) for untrained items, and 94.6% ($SD = 22.6\%$) for trained items. Accuracy for the sleep group was: 94.9% ($SD = 21.9\%$) for untrained items, and 96.4% ($SD = 18.6\%$) for trained items.
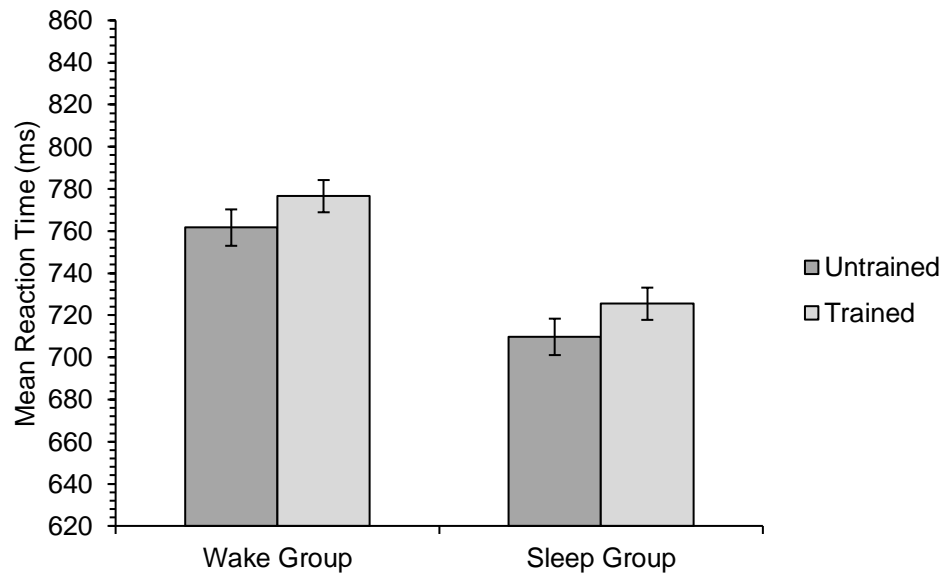
Figure 19. Experiment 5. Mean reaction time on the semantic relatedness judgement test for participants in each of the two groups for untrained and trained stimulus items. The data shown are for correct related trials only (trials to which the participants correctly responded 'yes' that the target and probe were semantically related). Error bars show standard errors for subject means, corrected for the within-subjects factor of training condition (Cousineau, 2005).

There was a significant main effect of group [$\chi^2(1) = 3.90$, $p = .048$], whereby the sleep group were faster overall than the wake group. The main effect of training was non-significant [$\chi^2(1) = 1.60$, $p = .205$], and there was no significant interaction between group and training condition as had been predicted [$\chi^2(1) = 0.09$, $p = .769$]. There was also no significant effect of block position [$\chi^2(1) = 0.06$, $p = .810$], interaction between group and block position [$\chi^2(1) = 0.92$, $p = .338$], or interaction between training condition and block position [$\chi^2(1) = 0.06$, $p = .809$]. The 3-way interaction was also not significant [$\chi^2(1) = 0.23$, $p = .632$].

Exploratory analysis: Items correctly recalled

For the purposes of obtaining a full and clear impression of the overall reaction time data from the semantic relatedness judgement task in order to inform future experiments, an additional exploratory analysis of the data was carried out. This analysis was of the subset of trials for which participants had correctly recalled the item in the cued recall test. This type of analysis was carried out by Rodd et al. (2012; Experiment 3), and was also appropriate for the present study to examine participants' implicit memory only for items where there was evidence that they had some explicit knowledge of the new meanings for the words.

The subset of the data for only items that had been correctly recalled by participants in the subsequent cued recall test can be seen in Figure 20. These data show a similar pattern to the overall data, with the sleep group responding faster than the wake group overall, and this time very little difference between the trained and untrained conditions, with no hint of any interaction.
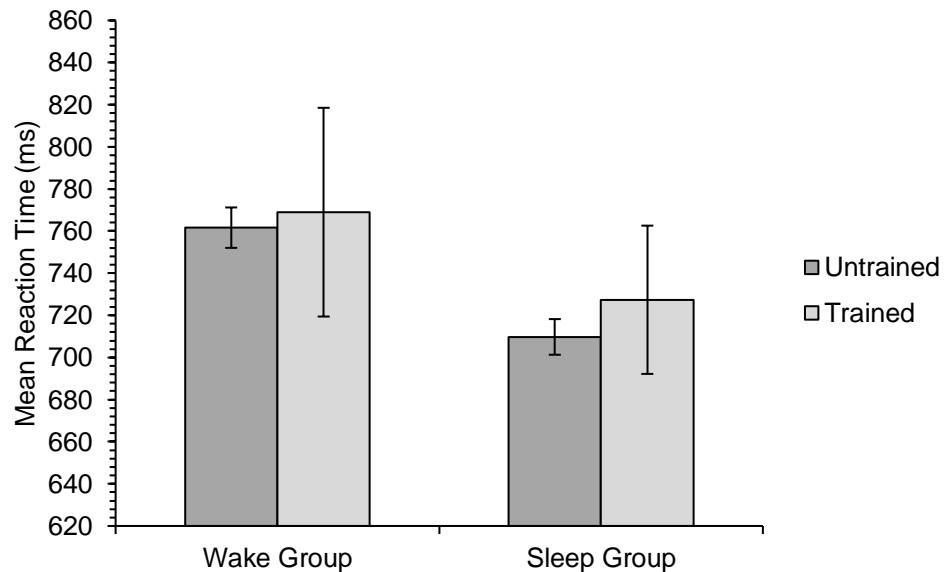
Figure 20. Experiment 5. Mean reaction time on the semantic relatedness judgement test for participants in each of the two groups for untrained and trained stimulus items only for the subset of trained items that participants correctly recalled in the cued recall test. The data shown are for correct related trials only (trials to which the participants correctly responded 'yes' that the target and probe were semantically related). Error bars show standard errors for subject means, corrected for the within-subjects factor of training condition (Cousineau, 2005).

For the analysis of this data, the first attempted model fit with the maximal model (which had exactly the same structure as the maximal model used in the analysis of the whole dataset) did not converge. The model was then simplified slightly (as recommended by Barr et al., 2013) by removing only the correlations between the random intercepts and slopes, after which the model converged and thus was used as the final model. Additionally, as for the main analysis the RT data were inverse-transformed in order to meet the assumptions of homoscedasticity and normality more closely and for ease of comparison with the whole dataset analysis. The likelihood ratio tests showed that there was a marginally non-significant main effect of group [$\chi^2(1) = 3.22$, $p = .073$], and no significant main effect of training [$\chi^2(1)$

$= 0.82$, $p = .366$] or training by sleep group interaction [$\chi^2(1) = 0.01$, $p = .903$]. The main effect of block position and all of the other interactions were also non-significant (all $p > .05$).

4.2.4 Discussion

The present study aimed to investigate whether sleep is important for integrating new word meanings into semantic memory, as has previously been shown for the learning of new spoken word forms (Davis & Gaskell, 2009). In the study participants were divided into two groups: participants in the sleep group were trained on new meanings for familiar words through story reading in the evening and tested the following morning, and participants in the wake group were trained in the morning and tested in the evening on the same day. There was 12 hours' delay between the training session and the test session, which consisted of an implicit measure of semantic relatedness judgement to assess consolidation of new meanings, as well as explicit memory measures of cued recall and multiple choice meaning-to-word matching. It was predicted that the sleep group would have better explicit memory of new word meanings, and only the sleep group would show interference from the new word meanings in the semantic relatedness judgement task that probed the pre-existing meanings of the words.

The results for the two explicit memory measures showed that participants in the sleep group remembered significantly more of the new word meanings than those in the wake group. Mean accuracy in cued recall of new meanings was 11.6% higher for the sleep group than the wake group, and accuracy on the multiple choice meaning-to-word matching task was 14.0% higher for the sleep group as compared to the wake group. These findings are consistent with those of Henderson et al. (2015) who found better cued recall of new word forms learned through stories at the 24-hour test than the immediate test, although performance in this earlier experiment could have been partially enhanced by a testing effect (as discussed in Chapter 3). Tamminen et al. (2010) also found significantly improved recall of new word forms for a group tested after 12 hours including sleep, but no improvement for a separate group of participants tested after 12 hours of wake. However, the present findings are in contrast to those of some other studies that have looked at the consolidation of new words and their meanings (Tamminen et al., 2012; Tamminen & Gaskell, 2013; Van der Ven et al., 2015). These studies found that explicit memory for new word meanings either remained the same (Tamminen et al., 2012), or decreased (Tamminen & Gaskell, 2013; Van der Ven et al., 2015) due to forgetting following a period of overnight sleep. The reason for the variation in the findings of these studies is unclear; this point will be addressed in detail alongside the findings of Experiment 6 in the general discussion of this chapter.

Participants in the sleep group responded significantly faster overall than those in the wake group on the semantic relatedness judgement task. This was not predicted, and could be due to circadian differences of the effects of time of day on encoding and/or test, as participants may simply be faster to respond in the morning (when the sleep group were tested), due to being more rested. Although no online measures of alertness such as the psychomotor vigilance task were used in the present study (Ashton, Jefferies, & Gaskell, 2018), there was no significant difference in participants' ratings on the Stanford sleepiness scale (Hoddes et al., 1973) between the sleep group and the wake group at the test session.

There was no significant overall effect of whether participants had been trained on new meanings for words in their response times on the semantic relatedness judgement task. Furthermore, the key interaction of interest between training and sleep group was also non-significant, which was contrary to the predictions. While it may be the case that insufficient consolidation had taken place to give rise to a semantic competition effect for the sleep group, this is somewhat inconsistent with the findings from the explicit memory measures. In the absence of any evidence of consolidation in the implicit memory measure, it is possible that the benefit of sleep for recall and recognition of the new meanings is due to passive protection against encoding of new information during the 12-hour period for the sleep group (Ellenbogen et al., 2006).

Another likely possibility is that the semantic relatedness judgement task lacked the sensitivity required to detect any consolidation effect. In particular, the length of time that the target word was presented onscreen and the inter-stimulus interval used in the task may have been too long. While this type of task has rarely been used in previous studies, two recent studies that have included a semantic relatedness judgement task have used much shorter durations for the presentation of the target word and inter-stimulus interval (Gilbert, Davis, Gareth Gaskell, & Rodd, 2018; Maciejewski et al., 2018). The long delay between the initial onset of the target word and the onset of the probe word is problematic in the present study, as it is likely that any potential disambiguation of the target word was fully resolved before the presentation of the probe word that related to the pre-existing meaning but not the newly-learned meaning. This issue is addressed in Experiment 6, where this task is adapted to make it more sensitive to measuring competition between newly-learned meanings and well-established meanings.

In summary, the results of Experiment 5 showed that participants had better explicit memory of new meanings for familiar words after 12 hours that included a period of overnight sleep, as compared with participants tested after a 12-hour period of wake. However, there was no difference between the sleep and wake groups in performance on the implicit measure

of consolidation of the new meanings. It is therefore unclear whether the benefit of sleep on explicit recall and recognition of the new meanings was due to active consolidation or due to passive protection from interference due to less opportunity for the sleep group to encode new information between training and test. However, the lack of evidence of competition arising between the novel and pre-existing meanings may have been due to a lack of sensitivity in the semantic relatedness measure, this task is therefore modified slightly for use in Experiment 6. Furthermore, it is possible that the sleep effect seen in recall and recognition of new word meanings in Experiment 5 may actually be an effect of time of day, as time of day was confounded with sleep group at both encoding and test. It is possible that participants may have learned better in the evening, or remembered better in the morning. This is an important consideration, and Experiment 6 goes some way to address this concern.

4.3 Experiment 6: Overnight consolidation after two study sessions

4.3.1 Introduction

This experiment was preregistered through the Open Science Framework; the preregistration can be retrieved from: https://osf.io/uvgp4 (Hulme & Rodd, 2017, August 9). Where applicable any deviations from the preregistration have been noted in the Method and Results sections for this experiment.

The previous experiment in this chapter investigated whether sleep is important for the consolidation of incidentally-learned new meanings for familiar words using a between-group manipulation for whether participants slept or remained awake in the 12 hours between training and test. Experiment 6 further investigates whether sleep is important for the consolidation of new meanings for already known word forms, and attempts to distinguish between active and passive benefits of sleep on memory of new word meanings. In Experiment 5 the benefit of overnight sleep for explicit memory of new word meanings could have been due to active consolidation or passive protection from interference. It was not possible to distinguish between these two accounts as participants were only trained in a single session, and no competition was found between the new and pre-existing meanings of the words that would have provided evidence of an active role for sleep in consolidating new word meanings. Furthermore, in Experiment 5 the sleep group manipulation was confounded with time of day for both encoding and test, also making it unclear whether the superior memory performance for the sleep group could be due to effects of sleep or time of day. Experiment 6 goes some

way to address these issues by way of a mixed "12:12" design somewhat similar to that used by Dumay and Gaskell (2007). In the present experiment participants are divided into two groups who are both trained twice, at two different times 12 hours apart, and who begin and complete the experiment at different times of day. Participants were trained and tested in two groups with two 12-hour delays between the first and second training sessions and the test session, with the AM test group beginning and completing the experiment in the morning, and the PM test group beginning and completing the experiment in the evening (see Figure 21). The two groups therefore have the same lengths of time delay between the two training sessions and the test (24 hours and 12 hours) and the same amounts of time spent asleep and awake, with the only difference being when the period of sleep occurs in relation to the test. At the test session, semantic integration of the new meanings for the words was assessed using an adapted version of the semantic relatedness judgement task from Experiment 5. The trial structure of the task was altered to match that used by Maciejewski et al. (2018), giving a much shorter delay between the onset of the target word and the onset of the probe word in order to increase the sensitivity of the task to participants' online semantic processing of the newly-ambiguous words. Finally, participants' explicit knowledge of the new word meanings was measured using cued recall and multiple choice meaning-to-word matching tests.
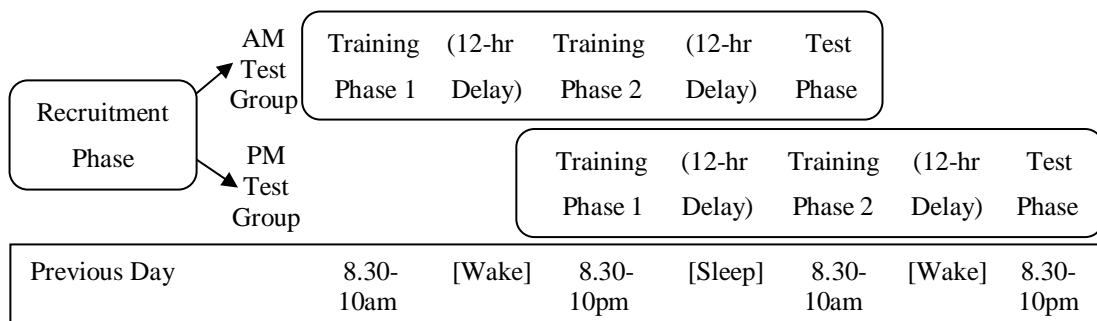
Figure 21. Diagram demonstrating the procedural design for the two groups in Experiment 6.

It is clear from Experiment 5 that new meanings for familiar words are not remembered as well following 12 hours spent awake as after 12 hours including overnight sleep. Gaskell et al. (2018) note that if the benefit of sleep is due to passive protection from interference, then 12 hours of interference before sleep would have the same detrimental effect as 12 hours of interference after sleep. This is because both groups would have spent a full day awake during which they encode new information that could interfere with the newly-learned word meanings. On the other hand, if sleep actively consolidates new word meanings, then items

learned shortly before sleep should have some protection against interference, which would not be the case for items learned immediately before a 12-hour period spent awake.

It was predicted that if sleep is important for the active consolidation of new word meanings then there would be better explicit memory for items that were learned immediately before overnight sleep than those that were learned immediately before a period of wake. Specifically, the PM test group would show better performance for items learned 24 hours ago than those learned 12 hours ago (as only items learned in the 24-hour delay session would have had the opportunity for sleep-based consolidation). The AM test group was predicted to show better performance for items trained in the 12-hour delay session than the 24-hour delay session, because items trained in the 12-hour delay session for the AM test group would be immediately followed by sleep. On the other hand, items trained in the morning 24 hours prior to test would have largely been forgotten during the 12 hours awake immediately following the training session, as shown in Experiment 5, leaving little to consolidate at the end of the day.

Importantly, the CLS account (Davis & Gaskell, 2009) would predict that if sleep plays an active role in consolidating new word meanings then the implicit memory measure should show inhibition for trained words due to competition between the new and old meanings of the words. This should only arise for items learned immediately before overnight sleep, and not for items trained before a 12-hour period of wake. The PM test group were therefore predicted to show a competition effect for items trained in the 24-hour delay session (slower reaction time compared with untrained items), and no competition effect for items trained in the 12-hour delay session. The AM test group were predicted to show a competition effect for items trained in the 12-hour delay session, but no competition effect for items trained in the 24-hour delay session

Alternatively, passive accounts of the benefit of sleep (Ellenbogen et al., 2006) would predict that sleep is only beneficial if it occurs both immediately after learning and immediately prior to test. Any intervening period of wake should cause interference due to the encoding of new information, and this interference should occur regardless of whether the period of wake occurs before sleep (as for the 24-hour delay condition for the AM test group) or after sleep (as for the 12-hour delay condition for the PM test group). In the present experiment, the period of overnight sleep immediately follows training and immediately precedes testing only for the 12-hour delay condition for the AM test group. It was therefore predicted that if the benefit of sleep is one of passive protection from interference, then only this condition would show better explicit memory for new meanings for familiar words, while all of the other conditions would show a similar lower level of recall and recognition accuracy.

Furthermore, a passive account of the role of sleep would not necessarily predict any competition effects between new and old meanings in the implicit memory measure for any of the conditions.

4.3.2 Method

Participants

Eighty-four adults participated and were included in the study (age: $M = 34.0$ years, $SD = 7.1$, range = 20-48), 69 of whom were female. All participants were monolingual native speakers of British English who had not been diagnosed with any reading or language disorders, and who had not taken part in any of the previous experiments in this thesis. None of the participants had been diagnosed with insomnia or any other sleep disorder or were currently taking any medication that could disrupt their sleep. Participants were recruited through the website Prolific Academic (Damer & Bradley, 2014) and were paid £10 in total for their participation upon completion of all sessions of the experiment.

In addition to the 84 participants included in the study, an additional 42 participants began the study but failed to complete all of the sessions and so no data was obtained for these participants. A further 25 participants were excluded due to getting more than one of the five comprehension questions wrong on any one of the stories they read. Four additional participants were excluded due to a technical error or attempting to complete a session of the study more than once. Upon inspection of the data, five participants were excluded due to being outliers in their mean reading speeds (faster than 657.2 words per minute, two standard deviations above the mean for all participants not already excluded for one of the aforementioned reasons). Finally, one participant was excluded due to having low accuracy on the semantic relatedness judgement test task (less than 75.4%, three standard deviations below the mean). The excluded participants were replaced with new participants to obtain the total of 84 participants included in the study.

Materials

Novel word meanings and short stories

The stimuli for the experiment were the same 16 words with novel semantically unrelated noun meanings as were used in the previous experiments (see Appendix A), which had been incorporated into the four separate short stories (see Appendix B). The same longer

paraphrased versions of the definitions as used in the previous experiments were again used in the multiple-choice meaning-to-word matching test in this experiment (see Appendix D).

Stimuli for semantic relatedness judgement test task

The stimuli for the semantic relatedness judgement task were the same as those used in Experiment 5, with an additional set of matched control words (see Appendix F for the full list). The experimental trials comprised of the 16 target words each paired with both a semantically related and a semantically unrelated probe word. Additionally, eight matched control words were also paired with both a related and unrelated probe (see Table 3). There were also 24 unmatched fillers, paired with either two semantically related or two semantically unrelated probes. Additionally, there were eight unmatched starting fillers each paired with two probes (with the same distribution of probe types as the experimental and matched filler trials) to serve as buffer trials at the start of each experimental block. Finally, another eight unmatched fillers were selected and each paired with two probes (with the same distribution of probe types as before) to serve as practice trials before the start of the experimental blocks.

	Trained Words ($n = 16$)		Untrained Control Words ($n = 8$)	
	Related Probes	Unrelated Probes	Related Probes	Unrelated Probes
Example	*dawn-dusk*	*dawn-basket*	*shield-sword*	*shield-baker*
No. of Letters	5.19 (1.17)	5.44 (1.15)	4.63 (1.06)	4.88 (0.64)
Frequency (per mil.)	26.28 (19.38)	17.59 (12.05)	20.16 (20.99)	17.23 (14.32)
Frequency (log-transf.)	3.58 (0.41)	3.45 (0.32)	3.43 (0.41)	3.37 (0.48)
WordNet Senses	5.56 (3.44)	5.50 (4.66)	3.50 (1.77)	4.13 (2.36)
Wordsmyth Senses	6.19 (4.05)	4.88 (2.75)	4.75 (2.38)	4.75 (3.45)
No. Semantic Associates	14.69 (5.26)	14.27 (3.69)[19]	11.25 (6.41)	12.00 (1.63)[20]
Target-Probe Relatedness	0.44 (0.18)	0.06 (0.07)	0.50 (0.15)	0.08 (0.07)

Table 3. Descriptive statistics for the lexical and semantic properties of the probe words used in the semantic relatedness judgement task in Experiment 6. The means for each measure are displayed in the table, with standard deviations given in parentheses. The words frequency data reported are the SUBTLEX-UK word frequencies in occurrences per million and log-transformations of the raw word

---

[19] There was no data for one item (*alien*, which was the semantically unrelated probe for *cake*).

[20] There was no data for one item (*basil*, which was the semantically unrelated probe for *barber*).

frequencies (log10[raw frequency+1]) (Van Heuven et al., 2014). Word sense data were taken from the WordNet (Fellbaum, 1998) and Wordsmyth (Parks et al., 1998) dictionaries. The number of semantic associates counts come from Nelson et al. (2004). The target-probe semantic relatedness values are Latent Semantic Analysis (LSA) estimates (Landauer et al., 1998).

Trials for the experimental task were split into two blocks of 56 trials each (eight starting fillers, followed by 16 experimental trials and eight control trials, intermixed with 24 filler trials), with a brief break between the two blocks. Each target word (stimulus, control, or filler) appeared once in each of the two blocks, paired with one of its two probe words in the first block, and the other in the second block. Half of the stimuli/control words (selected at random) appeared with their semantically related probe in block 1 and then with their semantically unrelated probe in block 2; the other half appeared with their semantically unrelated probe in block 1, then with their semantically related probe in block 2. The order of the two blocks was counterbalanced across participants. The order of trials within each of the two experimental blocks (and the starting fillers at the beginning of each of the two blocks) was randomised for each participant.

Design

About half of the participants ($N = 43$) were randomly assigned to the AM test group, and the other half ($N = 41$) were assigned to the PM test group. The pair of stories (stories 1 and 4, or 2 and 3) trained in the first or second training session was counterbalanced across participants. Finally, the order of the two blocks of trials in the semantic relatedness judgement task was counterbalanced across participants in order to minimise any potential order or repetition effects of seeing each stimulus twice (once with a semantically related probe, and once with an semantically unrelated probe). Participants were randomly assigned to one of these eight versions in the recruitment phase of the experiment.

Procedure

The experiment was carried out online using Qualtrics (Qualtrics, 2015) for the recruitment and training phases, and Gorilla (Gorilla.sc, 2017) for the testing phase. Figure 21 shows a schematic of the experiment, with the timescale for the training and testing of the two experimental groups.

Recruitment phase

As for Experiment 5, the study began with a recruitment phase in which participants were asked to provide some demographics details and to commit to taking part in all sessions of the experiment. The participants were at this point randomly assigned to one of the eight versions of the experiment, which determined whether they were assigned to the AM test group or the PM test group. Participants were given the times for their three subsequent sessions beginning the following day, for the AM test group these were: 8.30-10am, 8.30-10pm, and 8.30-10am the following day; for the PM test group these were: 8.30-10pm, and 8.30-10am and 8.30-10pm the following day (see Figure 21). Participants were not informed that the purpose of the study was to examine the learning of new word meanings, and were not aware that their memory would be tested. Instead they were told a cover story that the purpose of the study was to investigate reading ability and comprehension of texts at different times of day.

Training phase

The training phase consisted of two separate sessions spaced 12 hours apart. At the beginning of each of the two training sessions and the test session participants were asked to rate their alertness on the Stanford Sleepiness Scale (Hoddes et al., 1973). During each training session, participants read two of the short stories (either stories 1 and 4, or stories 2 and 3). The procedure for reading the stories and answering the simple multiple-choice comprehension questions was the same as for the previous experiments (see Chapter 2 for details). As with all of the previous experiments, the purpose of the story comprehension questions was to check that the participants had read the story carefully and fully processed the meaning, and therefore served as exclusion criteria. Participants were excluded if they got more than one of the five comprehension questions wrong on any one of the four stories they read: 25 participants were excluded for this reason. In each session, after completing the first story participants answered some questions about their enjoyment and clarity of the story they had just read, and a few questions about their reading habits in general (taking approximately 30 seconds in total) before they were allowed to continue and begin reading the second story. After reading the second story in each session participants were asked the same questions about their enjoyment and the clarity of the second story. The purpose of these questions was to maintain the pretence of the cover story that purpose of the experiment was to investigate reading ability and comprehension of texts at different times of day.

Testing phase

Semantic relatedness judgement task

As for Experiment 5, the first task of the testing phase was the semantic relatedness judgement task. Participants were presented with trials for stimuli and filler items one at a time in the centre of the screen. The timing of the trials for this task was much faster than for Experiment 5. Each individual trial began with a fixation cross presented for 500ms, followed by a brief blank screen for 100ms, the target word then appeared onscreen for 200ms, followed by another brief blank screen for 50ms, followed by the probe word which was presented until a response was given (with a time-out after 2000ms). Participants' task was to decide whether the target and probe word were semantically related (e.g., *hive-honey*) or not (e.g., *hive-bicycle*). Participants were not told which meaning of the target word they should attend to, and were instructed to try to respond as quickly and accurately as possible. They indicated their choice with a "yes" response (by pressing the "j" key), or a "no" response (by pressing the "f" key). If a response was not given until after the probe had been onscreen for 1500ms, then a message was displayed to tell the participant that their response was too slow and that they should respond more quickly.

Before beginning the experimental task, participants first completed a practice block of 16 trials. The purpose of the practice block was to familiarise participants with the task and to provide feedback on their speed and accuracy. Following each practice trial, a feedback screen informed them whether their response had been correct or incorrect. Additionally, if a response was slower than 1500ms, another feedback message told participants that their response was too slow. Following the practice block, participants proceeded to the first of the two experimental blocks, which each began with eight starting fillers in order to accustom participants to the speed and rhythm of the task. Following this there were 16 experimental trials and 16 matched filler trials in each of the two blocks, the order of which was randomised separately for each participant. There was a break of at least ten seconds between the two experimental blocks.

Cued recall test

Immediately following the semantic relatedness judgement task, participants were given a cued recall test for all of the new meanings for the words that they had been trained on through the stories. The procedure for this test was the same as for the previous experiments. Participants were presented one at a time with the 16 stimulus words for which they had been trained on a novel meaning; they were instructed to recall the appropriate new meaning for each of the words that they had encountered in the stories and type it into a blank text box.

They were asked to give as much detail as they could and to try to answer in full sentences even if they were unsure of their answer. If they could not remember anything about the new meaning for the words, then they were instructed to type "don't know". The order of presentation of the word cues was randomised for each participant.

Multiple choice meaning-to-word matching test

In the final test task, as in previous experiments, participants were presented one at a time with short sentences giving definitions of the 16 novel word meanings that they had been trained on through the stories. The sentences omitted the words to which they were referring, and for each novel meaning participants were asked to select the word that they thought matched the definition from a list of all eight of the stimulus words for which they had been trained on a new meaning in the same session. The foil items in each instance were the other seven words that participants had encountered new meanings for during the same training session. The order of the eight words to choose from was randomised for each test item, as was the order of presentation of the 16 new meaning definitions.

4.3.3 Results

Stanford Sleepiness Scale

The results for the Stanford Sleepiness Scale (SSS; Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973) measure taken at the beginning of each session of the experiment were analysed using three separate Wilcoxon-Mann-Whitney tests (one for each session of the experiment). The analysis showed that SSS scores at the first training session did not differ between the AM test group for whom this session was in the morning (*Mdn* = 3) and the PM test group for whom this session was in the evening (*Mdn* = 3), W = 728.5, *p* = .150. The SSS scores for the second training session also did not differ between the AM test group for whom this session was in the evening (*Mdn* = 3) and the PM test group for whom this session was in the morning (*Mdn* = 2), W = 1038.5, *p* = .149. Finally, SSS scores at the test session did not differ between the AM test group for whom this session was in the morning (*Mdn* = 3) and the PM test group for whom this session was in the evening (*Mdn* = 3), W = 895.5, *p* = .746.

Analysis procedure

Responses for the two explicit measures of cued recall and multiple choice meaning-to-word matching were coded in the same way as for previous experiments as either "1" for correct or "0" for incorrect. The reaction time (RT) data from the semantic relatedness judgement task were pre-processed as for Experiment 5 prior to analysis. RTs for incorrect trials were removed (5.5% of all experimental trials), and RTs were trimmed out if they were faster than 300ms or slower than 2500ms (0.1% of remaining experimental trials). The analysis for the semantic relatedness judgement task was therefore carried out on correct, related trials as for Experiment 5. The data were analysed using linear mixed effects (LME) models with the lme4 package (version 1.1-15; Bates et al., 2015) and R statistical software (version 3.3.3; R Core Team, 2017). Three models were created to separately analyse the results of the two explicit measures of cued recall and multiple choice meaning-to-word matching, and the implicit measure of reaction time on the semantic relatedness judgement task.

The logistic LME models used to analyse the accuracy data for the two explicit measures contained random effects by participants and items (with a slope by participants for training condition; and slopes by items for group, training condition, and the interaction between these variables). The logistic LME models for the explicit measures also contained fixed effects for group (two levels: AM test group or PM test group), training condition (two levels: 12 hours ago or 24 hours ago), and the interaction (which was created by multiplying the contrasts for these two variables). The contrasts were defined using deviation coding for group (AM test group: -0.5 vs. PM test group: 0.5), and the training condition (12 hours ago: -0.5 vs. 24 hours ago: 0.5).

The LME model used to analyse the RT data for the implicit measure also contained random effects by participants and items (with slopes by participants for the two contrasts for training condition; and slopes by items for group, one of the contrasts for training condition, and the interaction between these variables). The model also contained fixed effects for group (two levels: AM test group or PM test group), training condition (three levels: untrained, 12 hours ago, or 24 hours ago), and the interaction (created by multiplying the contrasts for these two variables). The contrasts were defined using deviation coding for group (AM test group: -0.5 vs. PM test group: 0.5), and two Helmert-coded contrasts for the fixed effect of training condition: one comparing the untrained condition with the two trained conditions combined (untrained: 0.67 vs. 12 hours ago: -0.33 vs. 24 hours ago: -0.33), and one comparing the 24-hour delay session with the 12-hour delay session (untrained: 0 vs. 12 hours ago: -0.5 vs. 24 hours ago: 0.5).

As for the previous experiments, the first attempted model fit in each case used the maximal random effects structure justified by the experimental design (as recommended by Barret al., 2013). For the two logistic LME models the optimiser was changed to "bobyqa" as recommended by Bates et al. (2016) for dealing with model convergence issues. The model for the multiple choice meaning-to-word matching measure converged with maximal random effects and was therefore used as the final model for the analysis, but the model for the cued recall measure did not converge. While the model for raw RT data from the semantic relatedness judgement task did converge with the maximal random effects structure, the assumptions of homoscedasticity and normality were violated in the raw RT data (see Appendices I and J), so the data were $\log_{10}$- and inverse-transformed (invRT=1000/rawRT) then modelled to compare with the raw RT data. As for Experiment 5, the inverse-transformed RTs most closely met the assumptions of homoscedasticity and normality and were therefore used for the analysis, however the model for the inverse-transformed RT data did not converge with the maximal random effects structure. The full models for the cued recall and semantic relatedness data therefore had to be simplified (following Barr et al., 2013) by removing the correlations between the random slopes and random intercepts for the random effects by participants and items (without removing any of the random slopes themselves). This allowed the full models for the cued recall and semantic relatedness measures to converge.

Significance of the main effects and interactions were assessed using likelihood ratio tests that compared the full models to identical models with only the factor or interaction of interest removed in turn (but leaving in any other interaction or main effect involving that factor or interaction, and always leaving the random effects structure intact).

Following on from the main analysis, planned simple effects analyses were carried out for the two explicit memory measures to determine whether there was a significant effect of training condition within either of the two groups. This was done by taking separate subsets of the data for the AM test group and PM test group and creating a model for each containing only a fixed effect for training condition (and random effects, with a slope by participants and by items for training condition). For the implicit semantic relatedness measure, planned follow-up analyses were carried out to determine (1) whether there was a significant interaction between group and any of the three pairs of levels of training condition, and (2) six simple effects analyses were carried out to determine whether there was a significant difference between any of the three pairs of levels of training condition within each of the two groups separately. Significance in the follow-up analyses of the interactions and simple effects was determined in the same way as for the main analyses by using likelihood ratio tests comparing the models containing the factor or interaction of interest to one without (while retaining the random effects structure; models for the simple effects analyses for the cued

recall and semantic relatedness judgement tests had the random correlations removed to match with the full models used in the main analysis for these measures).

Cued recall of novel meanings

The mean percentage accuracy data for the cued recall test are shown in Figure 22. Cued recall performance was very low overall and did not appear to differ by much between the 24-hour delay (16.5%) and the 12-hour delay (17.4%) for the PM test group. However, accuracy appeared to be slightly greater for the 12-hour delay (that is items that were trained in the second training session in the evening just prior to sleep; 27.0%) than the 24-hour delay (14.8%) for the AM test group.
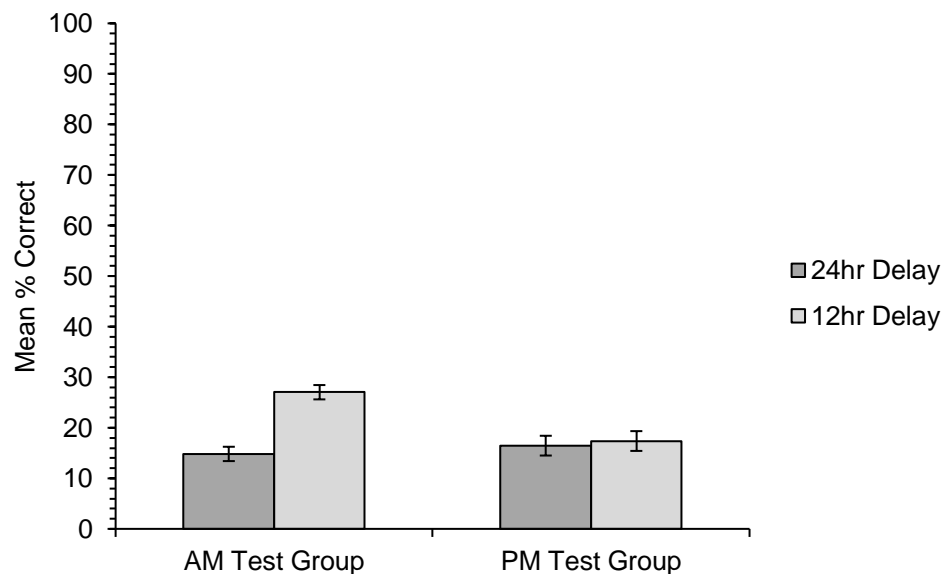
Figure 22. Experiment 6. Mean percentage of correct responses given on the cued recall test (meanings correctly recalled for the appropriate word) by participants in each of the two groups for new meanings for familiar words trained either 24 hours or 12 hours prior to test. Error bars show standard error for the subject means corrected for the within-subjects factor of training condition (Cousineau, 2005).

The analysis showed no significant main effect of group [$\chi^2(1) = 1.47$, $p = .226$], a significant main effect of training condition [$\chi^2(1) = 5.10$, $p = .024$], and a significant interaction between group and training condition [$\chi^2(1) = 4.74$, $p = .029$]. The planned simple effects follow-up analysis showed a significant difference between the training conditions for the AM test group [$\chi^2(1) = 11.45$, $p < .001$], but not for the PM test group [$\chi^2(1) = 0.10$, $p =$

.750]. (The p-values for these simple effects analyses were compared against a Bonferroni-corrected α of .025.)

Multiple choice meaning-to-word matching

The mean percentage accuracy data for the multiple choice meaning-to-word matching test are given in Figure 23. Accuracy on this task was much higher than for the cued recall test, although the pattern of the means was similar. There appeared to be little difference between the 24-hour delay (52.7%) and the 12-hour delay (53.7%) for the PM test group, and slightly higher accuracy for the 12-hour delay (56.7%) than the 24-hour delay 46.8%) for the AM test group.
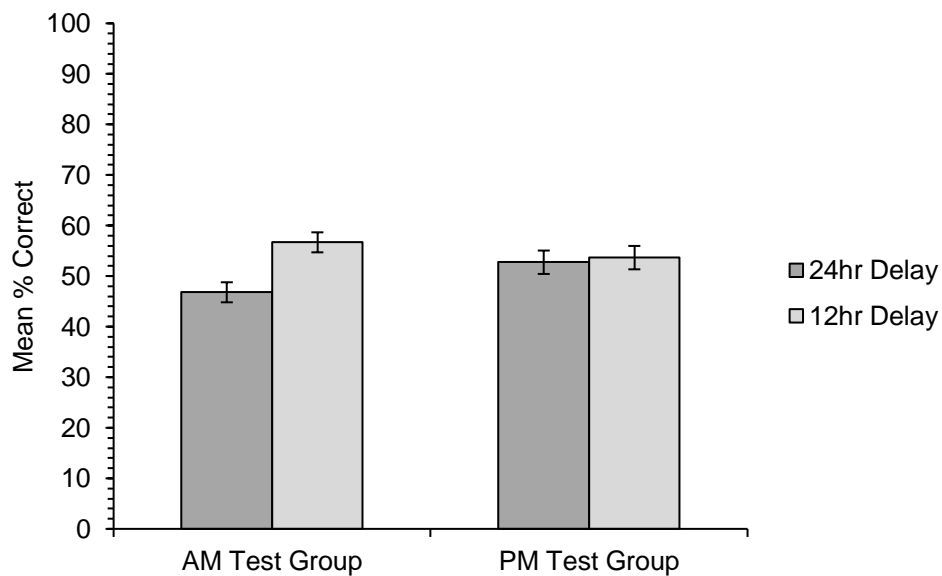
Figure 23. Experiment 6. Mean percentage of correct responses given on the multiple choice meaning-to-word matching test (words correctly paired with the appropriate meaning definition) by participants in each of the two groups for new meanings for familiar words trained either 24 hours or 12 hours prior to test. Error bars show standard error for the subject means corrected for the within-subjects factor of training condition (Cousineau, 2005).

The analysis for this measure showed no significant main effect of group [$\chi^2(1) = 0.07$, $p = .792$], as well as no significant main effect of training condition [$\chi^2(1) = 2.63$, $p = .105$], and no significant interaction [$\chi^2(1) = 2.46$, $p = .117$]. The planned simple effects follow-up analysis showed a significant effect of training condition for the AM test group [$\chi^2(1) = 5.29$,

$p = .021$], but not for the PM test group [$\chi^2(1) = 0.00$, $p = .964$] (with p-values for these simple effects analyses compared against a Bonferroni-corrected α of .025).

Semantic relatedness judgement

Participants' accuracy in identifying target words and probes as semantically related was high overall ($M = 91.9\%$, $SD = 7.5\%$), and only varied by a maximum of 2.9% between any of the means for the different conditions[21]; the accuracy data were therefore not analysed. The mean raw reaction time data for the semantic relatedness judgement task are given in Figure 24.
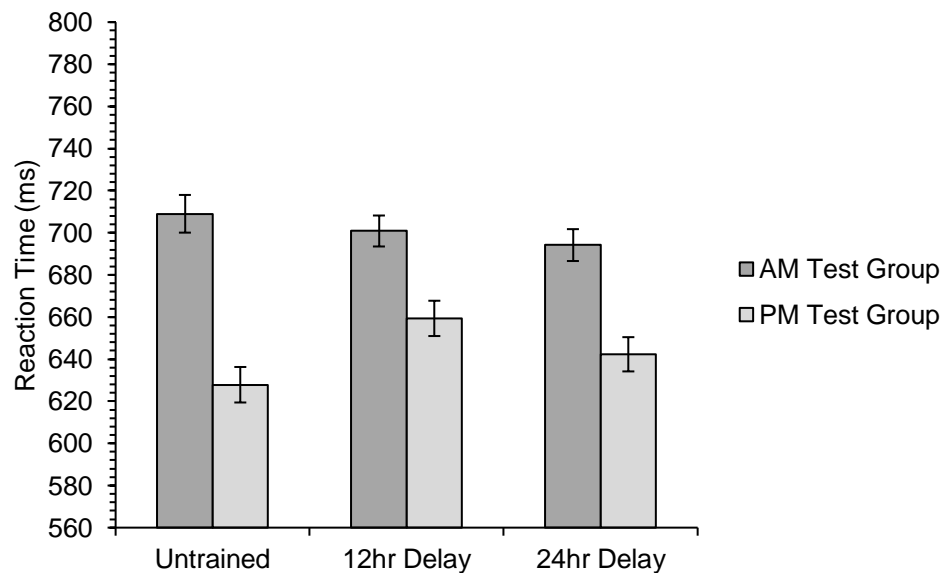
Figure 24. Experiment 6. Mean raw reaction time on the semantic relatedness judgement task for participants in the AM test group and PM test group on items that were either untrained, trained 12 hours prior to test, or trained 24 hours prior to test. The data shown are for correct related trials only (trials to which the participants correctly responded that the target and probe were semantically related). Error bars show standard errors for subject means, corrected for the within-subjects factor of training condition (Cousineau, 2005).

---

[21] The mean percentage accuracy data on the semantic relatedness judgement task for the two different groups were as follows. For the a.m. test group: untrained 93.9% ($SD = 24.0\%$), 12-hour delay 91.9% ($SD = 27.4\%$), 24-hour delay 91.0% ($SD = 28.7\%$). For the p.m. test group: untrained 92.4% ($SD = 26.6\%$), 12-hour delay 91.2% ($SD = 28.4\%$), 24-hour delay 91.2% ($SD = 28.4\%$).

The analysis showed that there was a significant main effect of group [$\chi^2(1) = 5.08$, $p = .024$] as the PM test group were faster overall than the AM test group. There was no significant main effect of training condition [$\chi^2(2) = 1.88$, $p = .391$], and the interaction was non-significant [$\chi^2(2) = 4.62$, $p = .099$]. The planned follow-up pairwise analysis of the interaction between group and the untrained and 12-hour delay levels of training condition was significant at the non-corrected level of $\alpha = .05$ [$\chi^2(1) = 3.90$, $p = .048$], however it was not significant at the Bonferroni-corrected level of $\alpha = .017$. The interactions between group and the untrained and 24-hour delay levels of training condition [$\chi^2(1) = 2.27$, $p = .132$], and between group and the 12-hour delay and 24-hour delay levels of training condition [$\chi^2(1) = 0.19$, $p = .665$] were also non-significant. The planned follow-up simple effects analyses for the 2x2 comparisons of training condition showed that for the AM test group there was no significant effect for untrained compared with 12-hour delay [$\chi^2(1) = 0.0001$, $p = .991$], untrained compared with 24-hour delay [$\chi^2(1) = 0.07$, $p = .791$], nor the 12-hour compared with the 24-hour delay [$\chi^2(1) = 0.45$, $p = .504$]. For the PM test group the comparison between the untrained and 12-hour delay was significant at the non-corrected level of $\alpha = .05$ [$\chi^2(1) = 3.90$, $p = .048$], but not at the Bonferroni-corrected level of $\alpha = .008$. The comparison between untrained and 24-hour delay [$\chi^2(1) = 1.25$, $p = .264$], and the comparison between the 12-hour and 24-hour delays was also non-significant [$\chi^2(1) = 1.41$, $p = .235$].

Exploratory analysis: Items correctly recalled

As for Experiment 5, in order to gain a better understanding of the overall data, exploratory analyses were carried out on the subset of data for only the items that had been correctly recalled in the subsequent cued recall test (see Figure 25). These analyses therefore allowed for the investigation of participants' implicit memory only for items where there was evidence that they had some explicit knowledge of the new meanings for the words (Rodd et al., 2012). It must be noted, however, that due to the low number of trials correctly recalled, power for these analyses is severely reduced.
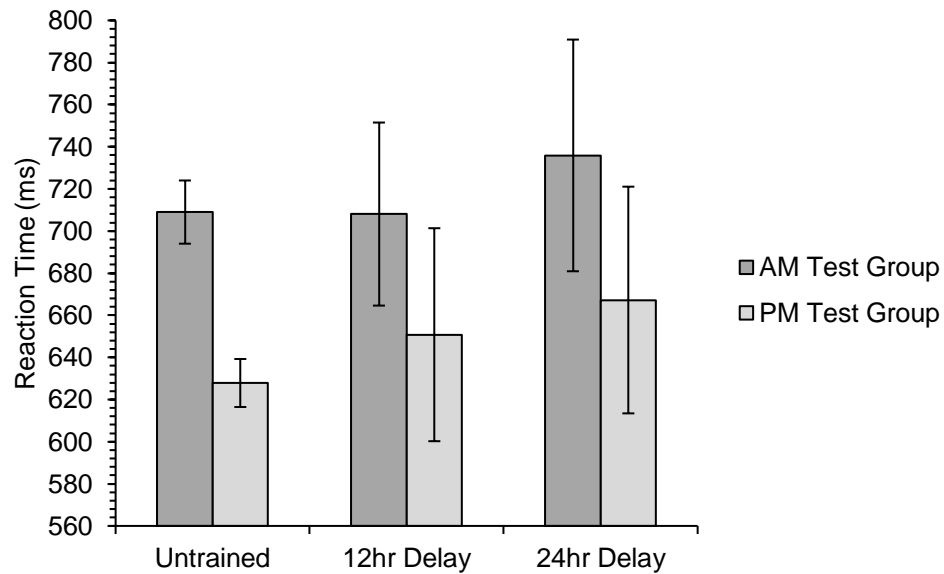
Figure 25. Experiment 6. Mean raw reaction time on the semantic relatedness judgement task for participants in the AM test group and PM test group on items that were either untrained, trained 12 hours prior to test, or trained 24 hours prior to test. Data shown are for the subset of trained items that participants correctly recalled in the cued recall test and for correct related trials only (trials to which the participants correctly responded that the target and probe were semantically related). Error bars show standard errors for subject means, corrected for the within-subjects factor of training condition (Cousineau, 2005).

There was a marginal but non-significant main effect of group [$\chi^2(1) = 2.89$, $p = .089$]. There was no significant main effect of training condition [$\chi^2(2) = 0.66$, $p = .720$], and the interaction was also non-significant [$\chi^2(2) = 1.04$, $p = .595$]. Follow-up pairwise analysis of the interaction between group and the untrained and 12-hour delay levels of training condition was non-significant [$\chi^2(1) = 0.54$, $p = .464$]. The interactions between group and the untrained and 24-hour delay levels of training condition [$\chi^2(1) = 0.89$, $p = .344$], and between group and the 12-hour delay and 24-hour delay levels of training condition [$\chi^2(1) = 0.02$, $p = .897$] were also non-significant (p-values were compared against a Bonferroni-corrected α of .017). Simple effects analyses for the 2x2 comparisons of training condition showed that for the AM test group there was no significant effect for untrained compared with 12-hour delay [$\chi^2(1) = 0.02$, $p = .896$], untrained compared with 24-hour delay [$\chi^2(1) = 0.004$, $p = .953$], nor the 12-hour compared with the 24-hour delay [$\chi^2(1) = 0.04$, $p = .844$]. For the PM test group the comparison between the untrained and 12-hour delay was non-significant [$\chi^2(1) = 0.59$, $p = .442$], as was the comparison between untrained and 24-hour delay [$\chi^2(1) = 1.37$, $p = .241$], and the comparison between the 12-hour and 24-hour delays [$\chi^2(1) = 0.05$, $p = .823$]. (The p-

values for the simple effects analyses were compared against a Bonferroni-corrected α of .008.)

4.3.4 Discussion

The aim of this study was to further examine whether sleep plays a role in the consolidation of new meanings for familiar words, and to try to tease apart active and passive accounts of the benefit of sleep for learning new word meanings. Experiment 6 built on the findings of Experiment 5 by dissociating sleep from encoding time to rule out time of day effects on learning of the new meanings. Experiment 6 also used a refined version of the implicit memory measure in order to improve sensitivity to detect any competition effects due to consolidation. The experiment took place over the course of 24 hours and used a mixed "12:12" design in which participants were trained twice and tested once in three separate sessions, with a 12-hour delay between each session. The AM test group began and completed the experiment in the mornings, and the PM test group began and completed the experiment in the evenings. Semantic integration of the new meanings was assessed using a speeded semantic relatedness judgement task to probe potential competition effects arising between the new and existing meanings of the words. Explicit knowledge of the new word meanings was assessed through cued recall and a multiple-choice meaning-to-word matching test. It was predicted that if sleep plays an active role in consolidating new word meanings, then participants would have better explicit memory for items that were learned prior to a period of sleep than those learned just before a period of wake, and consolidated items would be more resistant to interference during any subsequent period of wake. It was further predicted that there would only be evidence of consolidation in the implicit measure for items learned immediately before overnight sleep: the 24-hour delay condition for the PM test group, and the 12-hour delay condition for the AM test group. On the other hand, if the benefit of sleep is due to passive protection from interference, then participants would only have better explicit memory if sleep occurred in the immediate interval between training and test (only the 12-hour delay condition for the AM test group). Any intervening period of wake should cause interference from newly encoded information, regardless of whether the period of wake comes before or after overnight sleep. The passive account would also not predict any competition effects in the implicit measure for any of the conditions.

The two explicit tests showed a similar pattern of results. There was a significant interaction between group and training condition in the cued recall measure: the AM test group had significantly better recall of new meanings trained 12 hours prior to test (immediately preceding overnight sleep), than for items trained 24 hours prior to test, while there was no

such difference for the PM test group. The multiple-choice test showed a similar pattern of results, although the interaction was non-significant. These findings are in line with predictions of passive accounts of the benefit of sleep. The 24-hour delay condition for the PM test group did not show a higher level of recall or recognition accuracy, despite being immediately followed by sleep. This was presumably due to retroactive interference from the encoding of new information during the subsequent 12 hours spent awake the following day prior to test. A 12-hour period of wake therefore seemed to cause similar interference regardless of whether it occurred before sleep (as for the 24-hour delay condition for the AM test group) or after sleep, as was the case for items trained in the 24-hour delay condition for the PM test group. Only items trained in the 12-hour delay condition for the AM test group showed a benefit of overnight sleep on explicit recall and recognition, and this was the only condition in which sleep came immediately between training and test, with no period of wake. In this condition there was less opportunity for interference from irrelevant input in the intervening period between training and test, both preceding and following sleep, which may explain why sleep was beneficial in this case. This point will be discussed in greater detail alongside the findings from Experiment 5 in the general discussion of this chapter.

The results for the implicit measure showed that participants in the PM test group responded faster overall than those in the AM test group. While this is not of particular theoretical interest, this is in contrast to the findings of the previous experiment where the group tested in the morning had faster reaction times than those tested in the evening, while the opposite was the case in the present study. The reason for this difference is unclear; there was no significant difference in mean rating on the Stanford Sleepiness Scale between the two groups at either of the training sessions or the test session, as was also the case in Experiment 5. This difference may therefore not be explicable in terms of circadian differences due to the two groups being tested at different times of day. However, it is possible that participants' self-reported ratings on the SSS scale may not be sufficiently sensitive to their mental state, and tasks such as the psychomotor vigilance task (e.g., Ashton et al., 2018) may give a better indication of alertness in order to rule out the possibility of circadian effects at test.

For the semantic relatedness task, the key interaction between group and training condition to assess active consolidation of new meanings for familiar words was non-significant, and all three of the two-way interactions were also non-significant. It was predicted that if sleep plays an active role in consolidating new word meanings, then for new meanings learned before a period of sleep there would be slowed access to the existing meanings of the words (as compared with untrained control words) due to competition from the newly-consolidated novel meanings for the words, however no evidence was found for this. The same was the case for an analysis using only the subset of words for which

participants had correctly recalled the new meaning. The possible explanations for the lack of evidence for consolidation will be discussed in detail alongside the results of Experiment 5 in the general discussion in this chapter.

In sum, Experiment 6 showed that sleep seems to be beneficial for remembering new meanings for familiar words under certain conditions. Sleep seems to be specifically helpful for remembering new word meanings when it occurs in the immediate interval between training and test. In this respect, sleep may aid memory by passively protecting against interference from irrelevant knowledge that could lead to more forgetting. No evidence of active consolidation of the new meanings for the familiar word forms was observed in the present study, as was also the case for Experiment 5; possible reasons for this will be explored in the general discussion.

4.4 General discussion

The aim of the present chapter was to explore whether sleep is important for consolidating new meanings for familiar words, as has previously been found in many studies for the learning of novel word forms. The two experiments in this chapter used cued recall and multiple-choice tests to examine explicit knowledge of the new word meanings, and a speeded semantic relatedness judgement task to probe consolidation of the new meanings into semantic memory.

4.4.1 Explicit knowledge of new meanings for familiar words

In Experiment 5 participants had better explicit knowledge of new meanings for familiar words when training was followed by a 12-hour period including sleep as compared to a 12-hour period of wake. While this experiment showed a benefit of sleep for explicit knowledge of new word meanings, it was unclear whether this benefit was due to active consolidation (Davis & Gaskell, 2009), or due to passive protection against interference from newly-encoded information (Ellenbogen et al., 2006). In Experiment 6 recall of the new word meanings was only improved when participants were trained in the evening and tested 12 hours later the following morning, which was identical to the conditions under which sleep was found to benefit explicit memory of the new word meanings in Experiment 5. There was no benefit of sleep for items that were trained immediately prior to sleep and then tested 24 hours later following 12 hours of wake. These results are consistent with some of the existing

literature on sleep and word learning where overnight sleep has been seen to improve explicit memory for novel word forms (Dumay et al., 2005; Henderson et al., 2015; Takashima et al., 2014; Tamminen et al., 2010). However, the present findings contrast with those of previous studies of consolidation of novel words and their meanings (Tamminen et al., 2012; Tamminen & Gaskell, 2013; Van der Ven et al., 2015), in which explicit memory for word meanings remained unchanged or showed some decline following overnight sleep.

The pattern of results for the cued recall and multiple-choice measures in Experiment 6 most clearly point to a passive role for sleep in protecting memories for the new word meanings against interference from later encoding of information. While the 24-hour delay conditions for both groups had a later period of overnight sleep that occurred sometime between training and test, for the AM test group this training session was immediately followed by 12 hours awake, and the PM test group had a subsequent period awake in the 12 hours between their second training session and the test. Gaskell et al. (2018) note that if the benefit of sleep is due to passive protection from interference, then 12 hours of interference before sleep would have the same detrimental effect as 12 hours of interference after sleep. This is indeed what the pattern of results for Experiment 6 suggest. If the benefit of sleep were due to active consolidation, then there should have been better memory for items learned in the 24-hour delay session by the PM test group, because 12 hours awake after sleep should not affect more stable consolidated representations. In Experiment 6 the only condition that benefitted from sleep was the 12-hour delay condition for the AM test group, which was the only condition that had no period of wake either before, after, or in the absence of sleep.

The predictions of the CLS model for changes in explicit memory measures following sleep are somewhat unclear. Explicit memory measures are only intended to assess the lexical configuration of words, such as knowledge of their meanings (Leach & Samuel, 2007), which does not necessarily depend on consolidation. Lexical configuration has been shown to dissociate from the lexical engagement of words, that is their ability to interact with other items in the lexicon (Leach & Samuel, 2007), which is dependent on consolidation having occurred. In other words, explicit memory measures can test information that is stored in either episodic or semantic memory, without discriminating between the two. However, Davis and Gaskell (2009) note that delayed effects of consolidation are observable in some tests of explicit memory, provided that the task is sufficiently demanding. That is to say simple tasks used to probe explicit memory may be carried out solely with support from episodic memory. The easy tasks of definition selection used by Tamminen et al. (2012) and meaning recognition and meaning recall (with "known" or "unknown" responses) used by van der Ven et al. (2015) did not require participants to provide definitions of the new meanings and thus may not have tapped into semantic memory. On the other hand, the explicit memory tests of cued recall and

multiple choice meaning-to-word matching used in the present study were challenging for participants, and required participants to draw on information from across the varied exposures to the new meanings for words in the stories. It is therefore possible that the improvements in explicit memory of the new meanings following overnight sleep in Experiment 5 and Experiment 6 may be indicative of consolidation having taken place. However, improvements in performance on explicit memory measures alone cannot be taken as evidence for consolidation without corroborating evidence of lexical engagement provided by implicit memory measures.

4.4.2 Consolidation of new meanings for familiar words

Neither of the two experiments in this chapter showed evidence of any competition effects between the new meaning and the pre-existing meaning in the implicit measure of semantic relatedness judgement. There is therefore no evidence in the present studies of sleep playing an active role in consolidation of new meanings for familiar words. However, this null result is not necessarily consistent with the findings of a growing number of studies linking sleep directly to the consolidation of new word forms (e.g., Dumay et al., 2004; Tamminen et al., 2010), and new words and their meanings (e.g., Clay et al., 2007; Van der Ven et al., 2015). It is possible that the lack of competition effects in the SR task was due to the new meanings for the words not having been learned sufficiently well through the incidental learning procedure, as evidenced by the low performance overall on the explicit measures of word meanings knowledge in both Experiment 5 and Experiment 6. As shown in in Experiments 2 and 3 of this thesis, new meanings for familiar words are acquired more efficiently under intentional learning conditions. Previous studies in which new meanings for familiar words became integrated into semantic memory (Maciejewski et al., 2018; Rodd et al., 2012) used intentional learning paradigms and far more intensive training of new meanings for familiar words over the course of several days, as compared with the single incidental learning session used in the present study.

The speeded semantic relatedness judgement task is an appropriate implicit task to probe consolidation of new meanings through examining competition arising between the new and old meanings for the words. However, there were some challenges in implementing this measure in the present studies. Due to the use of the incidental learning paradigm in which participants encountered new meanings in naturalistic story contexts, the total number of new meanings participants could learn in a single session was limited to eight. This gave only a small number of items per participant in each condition and therefore lower power to detect

an effect, which combined with the relatively low explicit knowledge of the new meanings in both experiments means that the test likely lacked enough sensitivity.

The experiments in this chapter were designed to dissociate sleep from the simple passage of time, which some previous studies have failed to do (e.g., Van der Ven et al., 2015). However, in Experiment 5 sleep was confounded with time of day at both encoding and test. The group differences in overall speed on the SR task points to possible circadian differences, meaning that it is difficult to interpret whether the higher performance on the explicit memory tests for the sleep group are due to a sleep effect or effects of time of day. In Experiment 6 the sleep manipulation was separated from time of day at encoding, making it possible to judge the effects of 12-hour periods of sleep and wake on explicit memory of the new word meanings, but it was not possible to account for time of day at test for the two groups. Previous studies have attempted to dissociate sleep from time of day at test through the use of multiple test sessions at different times of day spaced 12 hours apart (Dumay & Gaskell, 2007). However, as highlighted in Chapter 3 of this thesis, such studies risk introducing a testing effect. Dissociating sleep from both the passage of time and the time of day at encoding and test remains a real challenge for sleep research. Future studies on the effect of sleep on consolidation of new word meanings could make use of polysomnographic recordings in combination with behavioural measures in order to probe the association between integration of new word meanings into semantic memory and specific components of sleep that have been linked to memory consolidation (Tamminen et al., 2010).

4.4.3 Conclusions

To conclude, this chapter examined whether sleep is important for the consolidation of new meanings for familiar words. The findings for the explicit memory measures in Experiment 5 and Experiment 6 are more consistent with passive accounts that explain the benefit of sleep in terms of protection against interference from the encoding of new information than the active role of sleep in consolidation as described by the CLS model of word learning. In both experiments sleep enhanced the recall and recognition of new meanings for familiar words, and Experiment 6 demonstrated that this benefit was specific to when overnight sleep occurred in the immediate interval between learning and test, without any intervening periods of wake. No evidence of competition effects between the new and pre-existing meanings of the words were found in the implicit semantic relatedness judgement measure in either of the two experiments, suggesting that the new meanings had not been sufficiently consolidated to engage with other items already in the mental lexicon.

Chapter 5: Concluding remarks

5.1 Theoretical contributions

The research reported in this thesis investigated adults' incidental learning of new meanings for familiar words. The aim of Chapter 2 (Experiment 1) was to examine the impact of the number of exposures on incidental learning of new meanings for familiar words from reading stories, and also to verify the newly-developed incidental learning paradigm. The findings showed that new meanings for familiar words can be learned from as little as two exposures in a story context, and accuracy in remembering new meanings increases in a linear trajectory with an increasing number of exposures. Furthermore, adults retain new meanings for familiar words well over the course of a week, regardless of the number of exposures during initial acquisition.

Chapter 3 compared incidental learning of new word meanings through story reading to acquisition under intentional learning conditions (Experiments 2 and 3). This chapter also investigated the effect of testing memory shortly after initial learning on the long-term retention of new meanings for familiar words (Experiments 3 and 4). Intentional learning was shown to be more efficient for acquisition, but new meanings learned incidentally through stories may be retained better over time. Retrieval practice at tests immediately after learning was found to be a powerful enhancer of future retention of new meanings for familiar words learned under either incidental or intentional learning conditions. Immediate cued recall or multiple-choice tests can be effective for this. These experiments highlight the potential benefits of combining incidental exposure to new vocabulary in texts with simple memory tests for language learners.

Finally, the experiments presented in Chapter 4 investigated whether consolidation during sleep plays a role in learning new word meanings, as predicted by the CLS model (Davis & Gaskell, 2009) based on evidence from spoken word form learning. Experiments 5 and 6 showed that sleep had a beneficial effect on explicit knowledge of new meanings for familiar words. However, as shown in Experiment 6, this benefit was specific to sleep occurring in the immediate interval between learning and test, without any long intervening period of wake. This suggests that sleep plays a passive role in protecting new memories of word meanings from interference by limiting the opportunity for encoding of any new information. However, no evidence was found for sleep playing an active role in consolidating new word meanings, as there was no indication of lexical engagement between the new meanings and existing meanings for words in either Experiment 5 or 6.

5.2 Methodological contributions

In addition to the theoretical contributions, the experiments in this thesis also highlight some important methodological considerations. Firstly, the research presented in this thesis is part of only a small amount of research that has focussed on adults' learning of new meanings for already-known words (Fang & Perfetti, 2017; Fang et al., 2016; Maciejewski et al., 2018; Rodd et al., 2012). Previous vocabulary learning studies that have investigated consolidation of word meanings have examined the learning of new word forms along with their associated semantic information (Clay et al., 2007; Tamminen et al., 2012; Van der Ven et al., 2015). Learning new meanings for familiar words differs in that attention is not divided between acquiring a word form and a meaning, but instead there is the need to reconcile a new meaning with existing representations for the meaning of that word. Research using new meanings for familiar words can provide an important contribution to this field, as it allows for the examination of semantic competition effects between the new and pre-existing meanings as a marker of consolidation, in a similar way as has been done previously with phonological competition between new and old word forms (e.g., Gaskell & Dumay, 2003). Although no evidence of such competition effects were found in the experiments reported in this thesis, this method may be useful for future studies trying to further tease apart the active and passive benefits of sleep for learning new word meanings.

A second contribution of the present research is that it highlights the possibility of using naturalistic texts in vocabulary learning research. The majority of studies on L1 vocabulary learning with adults have used designs based on explicit, intentional learning conditions. However, most new L1 words and their meanings are acquired incidentally through reading throughout late-childhood and adulthood. It is important to study vocabulary learning under more naturalistic conditions, as Experiments 2 and 3 of this thesis demonstrated that incidental and intentional learning conditions can have differential effects on acquisition and long-term retention of word meanings. Furthermore, the role of consolidation in vocabulary learning may differ depending on the learning conditions (Coutanche & Thompson-Schill, 2014; Fernandes, Kolinsky, & Ventura, 2009; Lindsay & Gaskell, 2013). The learning paradigm used in this thesis ideally combines naturalistic conditions for incidental vocabulary acquisition from reading with experimental control over exposure to items. This paradigm has the potential to be adapted for use in future studies to look at how a range of different factors might influence efficiency of learning and retention of new meanings for familiar words, such as attention, depth of processing, modality of story presentation, contextual diversity, or repetition of stories (e.g., M. Horst, 2005; Webb & Chang, 2015).

Thirdly, the findings of Chapter 3 provide an important reminder of the role that testing can play in the vocabulary learning process. This is of particular importance for word learning studies that compare participants' memory performance between multiple tests of the same material at different time points. As such, improvements in memory overnight that are explained in terms of offline consolidation during sleep may be at least partly due to a testing effect (e.g., Henderson et al., 2015). Experiments 3 and 4 demonstrated a clear benefit of retrieval practice on performance on later tests of explicit memory, and it is likely that prior testing may also have an effect on implicit memory measures used to probe consolidation of new words or meanings (Antony et al., 2017). Possible solutions to this dilemma for studies of overnight consolidation effects in vocabulary learning could be to test only some items in each test session (although testing some items may aid retention of untested items trained at the same time; Chan et al., 2006), or training participants twice and testing only once as in Experiments 5 and 6.

Finally, this thesis highlights the benefits of doing web-based research. All of the experiments in this thesis were run online: Experiments 1-6 used Qualtrics (Qualtrics, 2015), Experiment 5 additionally used Qualtrics Reaction Time Engine (Barnhoorn et al., 2015), and Experiment 6 used Gorilla (Gorilla.sc, 2017). A specific advantage for the experiments in this thesis has been the facilitation of multi-session studies. Experiments 1-4 included a surprise delayed test after one day (Experiments 2-4) or one week (Experiment 1); participants were invited to participate in the delayed test through a message without being given any prior notice. This was in order to discourage participants' use of intentional learning strategies due to expectation of a later test, and would have been difficult to achieve in a lab-based study. Furthermore, to the best of my knowledge, Experiments 5 and 6 in this thesis may be the first studies of the effects of overnight sleep on vocabulary learning to be run via the Internet. This gave substantial practical benefits in terms of being easier and less time-consuming than having participants come to the lab at specific times in the mornings and evenings, and allowed for multiple participants to be trained and tested simultaneously. Additionally, running the experiment online meant that participants were able to read the stories and be tested on their own computers in their own homes, improving the ecological validity of the research.

There are also several more general advantages of web-based research. Running studies online allows for quicker recruitment of participants, making large sample sizes much more attainable and therefore improving statistical power. Additionally, the use of online recruitment websites such as Prolific Academic (Damer & Bradley, 2014) allows access to a more diverse population, with the option to only recruit participants who meet specific demographics criteria. In contrast, many lab-based studies are limited to recruiting convenience samples of predominantly high SES (socioeconomic status) undergraduate

psychology students who are not representative of the wider population (Henrich, Heine, & Norenzayan, 2010). These considerations are therefore important to improve the replicability of psychological research.

However, there are a few commonly held concerns regarding web-based research. For example, one concern is how to ensure data quality and monitor activity of participants taking part in experiments remotely. I addressed this in the experiments in this thesis by examining participants' reading times for the stories along with their responses to easy comprehension questions. Another concern is how to confirm participants' eligibility. The Prolific Academic recruitment website (Damer & Bradley, 2014) asks participants in the pool to provide a wide range of demographics details that can be used by researchers to screen out participants who are not eligible to take part in a given study. While it is not ideal to rely on participants' self-reported demographics details to determine their eligibility, this problem is not unique to web-based research. A concern regarding the technical aspects of carrying out behavioural research online is whether reliable reaction time data can be collected from computers with varying Internet connection speeds (Plant, 2016). This is not an issue for most online experimental software (e.g., Gorilla; Gorilla.sc, 2017) because stimuli for all trials are pre-loaded prior to the start of an experiment, and reaction time data is recorded and stored locally on the computer before being uploaded upon completion. Furthermore, several recent studies have convincingly verified reaction time data collected online against equivalent data collected in the lab (Barnhoorn et al., 2015; Hilbig, 2016; McGraw, Tew, & Williams, 2000). Overall, the benefits of web-based research clearly outweigh the disadvantages, as it affords the opportunity to carry out certain experiments much more easily than can be done in the lab.

5.3 Future directions

The research reported in this thesis has highlighted some areas that warrant further investigation. For example, further work is needed to examine how new word meanings become integrated into semantic memory. The experiments in Chapter 4 found evidence of a passive benefit of sleep in protecting memories of new word meanings against interference from the encoding of new information, but no evidence was found for an active role for sleep in consolidating new meanings. However, previous studies have shown that new meanings for familiar words can compete with pre-existing meanings for access, therefore showing they become integrated into semantic memory with more extensive training over longer time periods (Maciejewski et al., 2018; Rodd et al., 2012). Further research is therefore required to tease apart active and passive benefits of sleep and to determine whether new word meanings

can become consolidated overnight with a more intensive training paradigm. Other experiments in this thesis suggest that an optimised training paradigm could include more exposures (Experiment 1), and intentional learning (Experiment 2).

Another question for future research to investigate is whether semantic relatedness between the new and old meaning of a word affects how a new meaning is consolidated. Rodd et al. (2012) previously showed that new meanings for familiar words are learned more easily when they are semantically related to the existing meaning as opposed to semantically unrelated. It is possible that novel semantically related meanings may be less reliant on offline consolidation than new unrelated meanings because they can connect more readily with existing representations for a word that are activated online during learning. A future study could therefore compare the emergence of semantic competition effects between novel related and unrelated meanings using a semantic relatedness judgement task as used in Experiments 5 and 6 and by Maciejewski et al. (2018).

Finally, an important unanswered question remains regarding the precise mechanism that underlies the testing effect. Experiments 3 and 4 showed a strong benefit of prior testing on long-term retention of new meanings for familiar words. Despite a growing amount of evidence that retrieval practice aids learning (for a review see Rowland, 2014), the underlying mechanisms remain unclear. An intriguing recent proposal is that retrieval may be a fast track to consolidation whereby the formation of hippocampal-neocortical representations is supported by online reactivation of related information (Antony et al., 2017). Further research is therefore required to determine whether the same neurocognitive mechanisms that are involved in offline consolidation are also responsible for the testing effect.

5.4 Conclusion

Vocabulary learning is highly important as it has long-term consequences for academic attainment and employment in later life. Native language vocabulary learning does not end in childhood but continues throughout the adult lifespan, and reading plays a particularly important part in this process. The research presented in this thesis has revealed some important characteristics of this learning process and provides the foundation for future research on this important topic. These studies have demonstrated the efficiency with which adults acquire new word meanings in their native language incidentally through reading and retain them well over time with assistance from sleep and retrieval practice.

References

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Ashton, J. E., Jefferies, E., & Gaskell, M. G. (2018). A role for consolidation in cross-modal category learning. *Neuropsychologia*, *108*, 50–60. https://doi.org/10.1016/j.neuropsychologia.2017.11.010

Azuma, T., & Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*(4), 484–504. https://doi.org/10.1006/jmla.1997.2502

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. https://doi.org/10.3758/s13428-014-0530-7

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2016). lme4: Linear mixed-effects models using "eigen" and s4. [Software Manual]. Retrieved from https://cran.r-project.org/web/packages/lme4/lme4.pdf

Batterink, L., & Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: An event-related potential study. *Journal of Cognitive Neuroscience*, *23*(11), 3181–3196. https://doi.org/10.1162/jocn_a_00013

Betts, H. N. (2018). Retuning lexical-semantic representations on the basis of recent experience. (PhD thesis). London, UK: University College London.

Bisson, M. J., Van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning*, *64*(4), 855–877. https://doi.org/10.1111/lang.12085

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A. F., Sommer, J., Wolbers, T., & Knecht, S. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *NeuroImage*, *25*(3), 958–968. https://doi.org/10.1016/j.neuroimage.2004.12.019

Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience*, *25*(5–6), 493–500. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18334767

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., & Rosen, B. R. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, *93*(25), 14878–14883. https://doi.org/10.1073/pnas.93.25.14878

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L. (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, *20*(7), 941–956. https://doi.org/10.1002/acp.1239

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73–101. https://doi.org/10.1016/j.cogpsych.2017.08.003

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities*, *44*(5), 431–443. https://doi.org/10.1177/0022219411410042

Cairney, S. A., Durrant, S. J., Hulleman, J., & Lewis, P. A. (2014). Targeted Memory Reactivation During Slow Wave Sleep Facilitates Emotional Memory Consolidation. *Sleep*, *37*(4), 701–707. https://doi.org/10.5665/sleep.3572

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319–343. https://doi.org/10.1017/S0305000904006749

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Clay, F., Bowers, J. S., Davis, C. J., & Hanley, D. A. (2007). Teaching adults new words: The role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 970–976. https://doi.org/10.1037/0278-7393.33.5.970

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General*, *143*(6), 2296–2303. https://doi.org/10.1037/xge0000020

Damer, E., & Bradley, P. (2014). Prolific Academic [Computer Software]. Retrieved from https://www.prolific.ac/

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word Learning: Homophony and the Distribution of Learning Exemplars. *Language Learning and Development*, *12*(3), 231–251. https://doi.org/10.1080/15475441.2015.1127163

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, *104*, 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, *21*(4), 803–820. https://doi.org/10.1162/jocn.2009.21059

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*, 3773–3800. https://doi.org/10.1098/rstb.2009.0111

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Oxford, UK: Blackwell.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice.

In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 31–344). San Diego, CA: Academic Press.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, *6*(3), 217–226. https://doi.org/10.1016/0361-476X(81)90002-3

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words: Research report. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Dumay, N., Gaskell, M. G., & Feng, X. (2005). A day in the life of a spoken word. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the Cognitive Science Society* (pp. 339–344). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from http://www.cogsci.northwestern.edu/cogsci2004/papers/paper554.pdf

Ellenbogen, J. M., Payne, J. D., & Stickgold, R. (2006). The role of sleep in declarative memory consolidation: passive, permissive, active or none? *Current Opinion in Neurobiology*, *16*(6), 716–722. https://doi.org/10.1016/j.conb.2006.10.006

Ellis, N. (1994). Consciousness in second language learning: Psychological perspectives on the role of conscious processes in vocabulary acquisition. In J. H. Hulstijn & R. Schmidt (Eds.), *Consciousness in second language learning AILA Review, 11* (pp. 37–57).

Fang, X., & Perfetti, C. A. (2017). Perturbation of old knowledge precedes integration of new knowledge. *Neuropsychologia*, *99*, 270–278. https://doi.org/10.1016/j.neuropsychologia.2017.03.015

Fang, X., Perfetti, C., & Stafura, J. (2016). Learning new meanings for known words: Biphasic effects of prior knowledge. *Language, Cognition and Neuroscience*, *32*(5), 637–649. https://doi.org/10.1080/23273798.2016.1252050

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output: Lexicalization during artificial language learning. *Cognition*, *112*(3), 349–366. https://doi.org/10.1016/j.cognition.2009.05.002

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*(4), 499–526. https://doi.org/10.1002/acp.1287

Gaskell, M. G., Cairney, S. A., & Rodd, J. M. (2018). Contextual priming of word meanings is stabilized over sleep. Retrieved from https://psyarxiv.com/52bxc/

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gaskell, M. G., & Ellis, A. W. (2009). Word learning and lexical development across the lifespan. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3607–3615. https://doi.org/10.1098/rstb.2009.0213

Gilbert, R. A., Davis, M. H., Gareth Gaskell, M., & Rodd, J. M. (2018). Listeners and readers generalize their experience with word meanings across modalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, (Advance online publication). https://doi.org/10.1037/xlm0000532

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

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*(1), 135–142. https://doi.org/10.1002/acp.2956

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182. https://doi.org/10.1016/j.jarmac.2014.05.003

Gorilla.sc. (2017). Gorilla.sc [Computer Software]. Retrieved from https://gorilla.sc/

Henderson, L. M., Devine, K., Weighall, A., & Gaskell, G. (2015). When the daffodat flew to the intergalactic zoo: Off-line consolidation is critical for word learning from stories. *Developmental Psychology*, *51*(3), 406–417. https://doi.org/10.1037/a0038786

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724.

https://doi.org/10.3758/s13428-015-0678-9

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, *10*(4), 431–436. https://doi.org/10.1111/j.1469-8986.1973.tb00801.x

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. https://doi.org/10.1016/S0022-5371(71)80029-4

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*, 1–11. https://doi.org/10.3389/fpsyg.2011.00017

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review*, *61*(3), 355–382. https://doi.org/10.3138/cmlr.61.3.355

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223. Retrieved from https://www.lextutor.ca/cv/beyond_a_clockwork_orange.html

Hulme, R. C., & Rodd, J. M. (2016). The testing effect in long-term retention of novel meanings for known words learned through incidental and intentional means. Retrieved from https://osf.io/e5zmk/

Hulme, R. C., & Rodd, J. M. (2017a). Learning new meanings for known words: Semantic integration and overnight consolidation. Retrieved from https://osf.io/uvgp4/

Hulme, R. C., & Rodd, J. M. (2017b). The impact of test method on the testing effect in long-term retention of new word meanings learned incidentally from stories. Retrieved from https://osf.io/c59tz/

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

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Malden, MA: Blackwell. https://doi.org/10.1002/9780470756492.ch12

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667. Retrieved from http://psych.wustl.edu/amcclab/Manuscripts/Jacoby/Jacoby (1978)/Retention Effects of Solving vs. Remembering.PDF

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*(4), 605–612. https://doi.org/10.2307/1414040

Kang, S. H. K., Mcdermott, K. B., Roediger, H. L., & Kang, S. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Kapnoula, E. C., & McMurray, B. (2015). Newly learned word forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin & Review*, *23*(2), 491–499. https://doi.org/10.3758/s13423-015-0897-1

Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*, 85–99. https://doi.org/10.1016/j.cognition.2014.09.007

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrival for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17–29. https://doi.org/10.1016/J.JML.2012.02.004

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*(3), 227–239. https://doi.org/10.1016/j.jml.2009.11.010

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In *The computer and literary studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

Konopak, B., Sheard, C., Longman, D., Lyman, B., Slaton, E., Atkinson, R., & Thames, D. (1987). Incidental versus intentional word learning from context. *Reading Psychology: An International Quarterly*, *8*(1), 7–21. https://doi.org/10.1080/0270271870080103

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, *55*(4), 306–353. https://doi.org/10.1016/j.cogpsych.2007.01.001

Lehmann, M. (2007). Is intentional or incidental vocabulary learning more effective? *The International Journal of Foreign Language Teaching*, *3*(1), 23–28.

Lindsay, S., & Gaskell, M. (2009). Spaced learning and the lexical integration of novel words. In N. A. Taatgen & H. Van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2517–2522). Amsterdam, NL: Cognitive Science Society. Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/575/paper575.pdf

Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(2), 608–622. https://doi.org/10.1037/a0029243

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9504270

Maciejewski, G., Rodd, J. M., Mon-Williams, M., & Klepousniotou, E. (2018). The cost of learning new meanings for familiar words. https://doi.org/10.17605/OSF.IO/7ydkw

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The

production effect: Delineation of a phenomenon. *Journal of Experimental Psychology Learning Memory and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin and Review*, *14*(2), 194–199. Retrieved from https://link.springer.com/content/pdf/10.3758/BF03194051.pdf

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Mason, B., & Krashen, S. (2004). Is form-focused vocabulary instruction worth while? *RELC Journal*, *35*(2), 179–185. https://doi.org/10.1177/003368820403500206

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. Retrieved from http://psycnet.apa.org/psycinfo/1995-42327-001

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(45), 494–513. https://doi.org/10.1080/09541440701326154

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The Integrity of Web-Delivered Experiments: Can You Trust the Data? *Psychological Science*, *11*(6), 502–506. https://doi.org/10.1111/1467-9280.00296

Mestres-Missé, A., Càmara, E., Rodriguez-Fornells, A., Rotte, M., & Münte, T. F. (2008). Functional neuroanatomy of meaning acquisition from context. *Journal of Cognitive Neuroscience*, *20*(12), 2153–2166. https://doi.org/10.1162/jocn.2008.20150

Mestres-Missé, A., Rodriguez-Fornells, A., & Münte, T. F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, *17*(8), 1858–1866. https://doi.org/10.1093/cercor/bhl094

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

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

Nation, P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language*, *27*(1), 136–145. Retrieved from http://nflrc.hawaii.edu/rfl/April2015/discussion/nation.pdf

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Ngo, H.-V. V., Martinetz, T., Born, J., & Mölle, M. (2013). Auditory Closed-Loop Stimulation of the Sleep Slow Oscillation Enhances Memory. *Neuron*, *78*(3), 545–553. https://doi.org/10.1016/j.neuron.2013.03.006

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Oppenheim, G. M. (2015). Competition in the expanding lexicon: Production reveals immediate semantic integration of newly acquired words. In *Architecture and Mechanisms of Language Processing* (p. 34). Valletta, Malta.

Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(6), 1543–1547. https://doi.org/10.1037/a0020604

Parks, R., Ray, J., & Bland, S. (1998). Wordsmyth English Dictionary-Thesaurus [Electronic version]. Chicago, IL: University of Chicago. Retrieved from https://www.wordsmyth.net

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading. *Studies in Second Language Acquisition*, *38*(1), 97–130. https://doi.org/10.1017/S0272263115000224

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, *22*(1), 31–55. Retrieved   from

https://pdfs.semanticscholar.org/5dfb/1e06263ac305633905efa2396ef01bdad573.pdf

Perfetti, C., Wlotko, E., & Hart, L. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1281–1292. https://doi.org/10.1037/0278-7393.31.6.1281

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, *59*(1), 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*(1), 408–411. https://doi.org/10.3758/s13428-015-0577-0

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Qualtrics. (2015). Qualtrics Survey Software. Provo, Utah, USA: Qualtrics. Retrieved from https://www.qualtrics.com

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Raven, J., Raven, J. C., & Court, J. H. (1998). Manual for Raven's progressive matrices and vocabulary scales. Section 5: The Mill Hill vocabulary scale. San Antonio, TX: Harcourt Assessment.

Reading habits. (2013). Retrieved from http://testyourvocab.com/blog/2013-05-09-Reading-habits

Rieder, A. (2003). Implicit and explicit learning in incidental vocabulary acquisition. *Vienna English Working Papers*, *12*(2), 24–39.

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*(7), 1095–1108. https://doi.org/10.3758/s13421-012-0209-1

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37. https://doi.org/10.1016/J.JML.2015.10.006

Rodd, J. M., Gaskell, 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. M., Gaskell, M. G., & Marslen-Wilson, W. D. (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., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). 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

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(1), 233–239. https://doi.org/10.1037/a0017678

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619. https://doi.org/10.1017/S0272263199004039

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Spivey, M., & Cardon, C. (2015). Methods for studying adult bilingualism. In J. W. Schwieter (Ed.), *The Cambridge handbook of bilingual processing* (pp. 108– 132). Cambridge, UK: Cambridge University Press.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/doi.org/10.1598/rrq.21.4.1

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32*(4), 827–853. https://doi.org/10.1017/S0305000905007099

Storkel, H. L., Maekawa, J., & Aschenbrenner, A. J. (2013). The effect of homonymy on learning correctly articulated versus misarticulated words. *Journal of Speech, Language, and Hearing Research*, *56*(2), 694–707. https://doi.org/10.1044/1092-4388(2012/12-0122)

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert vs. overt retrieval. *Frontiers in Psychology*, *8*, 1018. https://doi.org/10.3389/fpsyg.2017.01018

Takashima, A., Bakker, I., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Richness of information about novel words influences how episodic and semantic memory networks interact during lexicalization. *NeuroImage*, *84*, 265–278. https://doi.org/10.1016/j.neuroimage.2013.08.023

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112. https://doi.org/10.1016/j.cognition.2012.06.014

Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *Quarterly Journal of Experimental Psychology*, *61*(3), 361–371. https://doi.org/10.1080/17470210701634545

Tamminen, J., & Gaskell, M. G. (2013). Novel word integration in the mental lexicon: evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology (2006)*, *66*(5), 1001–1025. https://doi.org/10.1080/17470218.2012.724694

Tamminen, J., Lambon Ralph, M. a, & Lewis, P. a. (2013). The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *Journal of Neuroscience*, *33*(39), 15376–15381. https://doi.org/10.1523/jneurosci.5093-12.2013

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *Journal of Neuroscience*, *30*(43), 14356–14360. https://doi.org/10.1523/jneurosci.3028-10.2010

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*, 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Tran, R., Rohrer, D., & Pashler, H. (2014). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*(1), 135–140. https://doi.org/10.3758/s13423-014-0646-x

Van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage*, *78*, 94–102. https://doi.org/10.1016/j.neuroimage.2013.03.071

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

Van der Ven, F., Takashima, A., Segers, E., & Verhoeven, L. (2015). Learning word meanings: Overnight integration and study modality effects. *PLoS ONE*, *10*(5), e0124926. https://doi.org/10.1371/journal.pone.0124926

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

https://doi.org/10.1080/17470218.2013.850521

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163. Retrieved from http://www.robwaring.org/papers/various/waring_takaki.pdf

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245. https://doi.org/10.1177/1362168814559800

Webb, S., & Chang, A. C.-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Wiley, J., George, T., & Rayner, K. (2018). Baseball fans don't like lumpy batters: Influence of domain knowledge on the access of subordinate meanings. *Quarterly Journal of Experimental Psychology*, *71*(1), 93–102. https://doi.org/10.1080/17470218.2016.1251470

Williams, R., & Morris, R. (2010). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, *16*(1–2), 312–339. https://doi.org/10.1080/09541440340000196

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Appendices

Appendix A

Table I. List of stimulus words and definitions of their novel meanings.

Stimulus Word	Novel Meaning Definition
Story 1: *Pink Candy Dream*	
Hive	A new Chinese-made type of small car designed for inner-city living, with reduced boot space but extra storage in side pockets at the front of the car.
Vase	A colloquial term for a base used for criminal operations, they are chiefly used by big-city criminal gangs as places to meet in secret and carry out illegal dealings.
Path	The smallest surveillance device ever invented, it has a tiny camera through which it records and feeds back video, and is mobile and can be moved around by remote control.
Foam	A safe that is incorporated into a piece of furniture with a wooden panel concealing the key lock, and each is individually handcrafted so that no intruders are able to recognise the chief use of the furniture.
Story 2: *Prisons*	
Dawn	A biomedical implant fitted around a pacemaker to protect against electromagnetic interference, to which they are very susceptible, by acting as a barrier against electrical and magnetic signal.
Spy	The residual inner core left behind when a star dies, which are unique to each celestial body and can only be viewed through the world's most powerful telescopes.

Feast	A suit worn to protect against extremely high levels of harmful radiation, it covers the whole body with just a window to see through, but is particularly itchy and uncomfortable to wear.
Pearl	A new medical device which is attached to the body and can take and record measurements from the blood without piercing the skin that can be transmitted to a receiver in hospital.

Story 3: *Reflections upon a Tribe*

Bruise	A type of traditional folk band which is made up of all male members, and when a player retires, their closest living relative is expected to take over their position which is considered a great honour.
Fog	A type of dance dating back centuries that is mainly performed by street performers, it involves elongating the body and swaying from side to side whilst keeping the head still.
Cactus	A unique and valuable type of precious stone that is often used in jewellery, it changes colour in a matter of seconds depending on the temperature and humidity.
Carton	A folkloric monster that walks on its two hind legs and has a fixed, mischievous smile, and is said to eat livestock.

Story 4: *The Island and Elsewhere*

Rug	A traditional type of wooden fishing boat used by communities on some Pacific islands, it requires two people to operate it and can move at a fast pace when the sea is calm.
Rust	The name for a small village in a clearing of land in the middle of the forest in which the houses are close together and the surrounding trees provide good shelter.
Fee	The name for the flat top of the forest canopy which is thick with different trees and plants interwoven; islanders believe it is the sacred realm of their ancestral spirits.

Cake   A traditional tribal headdress decorated with feathers, shells and furs
              which is worn for religious ceremonies celebrating man's relationship
              with nature, the land and the sea.

Appendix B

The four stories that were used as the stimulus materials for all of the experiments in this thesis. The stimulus words have been underlined for transparency, but were not highlighted in any way when presented to participants in the experiments.

Story 1: Pink Candy Dream

The minute I pulled open the door and squeezed myself in behind the wheel I knew that Marla had borrowed the hive to take Lilly to playgroup again; the side pockets were bulging with plastic animals, baby wipes and half-eaten breadsticks. I extracted a neon pink sippy cup from the drinks holder and replaced it with my double shot Americano and wedged my kit bag into the hive's tiny passenger seat on top of a fleecy jacket and a sparkly pink slipper. Not for the first time, I was struck by a pang of longing for the old days - when cars came with luxuries like a boot, a back seat and more than three millimetres of leg room.

'Downtown! Quadrant Three!' I instructed the dashboard.

The electrics flickered on and the hive eased out into the flow of traffic. We climbed quickly to the top of the six road levels, squeezed between a pair of commuter buses and darted into a side street. As we zipped past nightclubs, betting palaces, sugar dens and bingo halls, I wondered how many of them had been taken over as vases; establishments used by the big criminal gangs for their clandestine deals and deliveries.

'Dammit!' I groaned, wiping a smear of mashed banana from my sleeve.

'Hey, Mike, you sound a little tense. How about we swing by the mall for a bubble tea? Wouldn't that be awesome?' The hive's digital voice was that of a relentlessly chirpy Californian teenager with a brand of American slang, which the car's Chinese designers must have picked up from old episodes of *The Simpsons*. 'Or what say a relaxing head massage?' the hive persisted.

'Just shut up and drive!' I snapped. When I was a kid, equipment was seen and not heard; it didn't pipe up with 'useful' suggestions every five minutes.

Don't get me wrong. I'm not usually one of those *everything-was-better-in-the-old-days* kind of a guy. Who'd want to go back to a time days before we had decent hair implants and a cure for dementia? And the hive could scurry through midtown traffic like nothing else. My bad temper was a stress thing, that's all. I'd been summoned for a "breakfast meeting" with

Control. And Zinnia Mendez wasn't the kind of boss who called her operatives in for chitchat and croissants.

<center>***</center>

<center>[SCREEN PAGE BREAK]</center>

 'How has this happened?' Controller Mendez demanded as I entered her corner office. She was staring out of the floor-to-ceiling window and didn't turn round. Instead she spoke her words to the pane of tinted smart glass and the muddy sky beyond. 'I *thought* we were keeping the Olafson gang under 24/7 watch?'

'We are,' I told the back of her jacket. The golden sheen suggested it was the latest spider silk microfibre. 'We've had nests of paths monitoring all their active vases for weeks,' I said, referring to the minute bugging devices that used the latest in nano-communications technology. Invisible to the naked eye, they were suspended in an adhesive solution and sprayed directly onto walls and ceilings.

Mendez didn't speak. A Control helicopter buzzed past the window.

Still standing, I gulped my cold coffee. The paths were my baby. It was like having thousands of tiny eyes and ears relaying messages back to Control. *Had something gone wrong?*

A very large number flashed before my eyes; the annual cost for the exclusive school we'd signed Lily up to. I couldn't afford to lose my job over this . . .

Mendez whipped round and leaned over her granite desk until her face was so close I could smell her lipgloss. *'They are one hundred percent undetectable!'* she mocked in a whiny voice, which seemed intended as an imitation of my own. *'Those Olafson scumballs won't suspect a thing . . .'*

'They are,' I spluttered. 'They won't.'

'So how do you explain *this?'* Mendez snarled. She pushed back from the desk and clicked her fingers. A virtual screen appeared between us and began to play surveillance footage.

The room was familiar. I'd installed the paths there myself. The walls were panelled in genuine antique pine and the floor carpeted in synthetic tiger skin. It was a lot swankier than the usual pool halls and strip clubs that the Olafson gang used as their 'business vases' for dropping off consignments of drugs and other illegal goods and picking up money. This was Olafson's personal penthouse apartment on the lagoon. It had been a nightmare to get in and out past security.

On screen a tall, thickly-bearded man entered the room. I recognized Olafson from a previous vase bust. He glanced around before flopping onto a yellow velvet sofa. He held up a small device and rotated it above his head.

'Is *that* what this is about?' I almost laughed with relief. It was common knowledge that all the gangs used portable scanners – nicknamed Rentokills - to make sure their vases were free of bugs and cameras. Obviously, I'd made sure that paths couldn't be picked up by any of the Rentokill models on the market.

Controller Mendez skewered me with a look. 'Just keep watching.'

[SCREEN PAGE BREAK]

Olafson was now peering at the armrest at the end of the sofa. It was one of those old Victorian types - with an arm at just one end – that are all the rage these days. *Chaises-longues,* I think they're called. Beneath the velvet cover, the armrest was polished wood, fashioned into an ornate scroll shape. As the camera feed zoomed in closer I couldn't help a moment of pride in my design – the paths use artificial intelligence to cluster into the best position and they work together to fine-tune the signal. Now they were magnifying the hairs on the back of Olafson's freckled hand to a forest. His fingers became those of a giant. A glint of silver revealed that they were curled around a delicate key. I held my breath as he slid the key into the centre of the scroll. A section sprang open and he pulled out a long, engraved tube.

'Yes,' I murmured, breathing again. 'I *knew* it! It's a foam**.**' The ornate *chaise longue* was no ordinary piece of furniture, but concealed a built-in safe with an intricate key-operated locking system.

I looked up at Zinnia Mendez. Her mouth was still knotted into a scowl. I couldn't figure out why she was so ticked off. 'Isn't this exactly what we've been looking for?' I asked. 'If we can find their foams we can find their paperwork . . . and bust them!'

'There's the small matter of *unlocking* the foams first, Michael!'

The voice came from near the door. I span round, my heart thumping in my chest. I hadn't heard anyone come in. An angular man in plain black uniform moved noiselessly towards the desk, his small feet seeming to slide rather than walk across the floor. Mendez registered his presence with a twitch of her eyebrow but didn't introduce us. I assumed he was from Internal Investigations. He narrowed his pale eyes in my direction. 'The foam's useless without knowing where Olafson keeps his keys.'

Now tell me something I don't know! I stopped myself saying the words out loud. You didn't want to get on the wrong side of Internal. I closed my eyes. Maybe I *am* one of those guys who thinks everything was better in the old days; it was so much easier when we just had to break computer passwords and decrypt files. But – like most of the serious crime rings - Olafon's gang had given up doing business on the internet. Hackers – on both sides of the law – had got so good at breaking down digital security that it was too dangerous. There was nowhere to hide. The gangs were all moving out of cyberspace, leaving it to the pornographers and petty scammers. The big money – drugs, weapons, tax fraud – had gone off-line.

Nowadays they kept their records on *paper.*

[SCREEN PAGE BREAK]

At first they'd stored their client lists and dodgy contracts in regular safes, but when those had become too easy a target, they'd moved on to less obvious hiding places. Foams were the latest trend – small cavities concealed in the frames of seemingly ordinary pieces of furniture. It had become something of a status symbol to commission bespoke cabinets, bureaus, love seats and *chaises longues* with the most cunningly hidden and exquisitely complex locking mechanisms. Skilled locksmiths could name their price. Hiding the tiny keys had become an art in itself; dentists had been known to hollow out teeth.

I turned my attention back to the screen. Olafson had unfurled a scroll of paper from the tube. He placed a little wooden clipboard on his lap, took a fat glossy fountain pen from his shirt pocket and began to write, his tongue sticking out from the corner of his mouth as he formed the letters. He blotted the ink with a cloth – he was clearly one of those gang bosses who'd taken retro writing to the extreme.

I was still trying to figure out why I was being hauled over the coals when Mendez clicked her fingers and told the film to move forward one hour.

Olafson had left the room. Workmen had arrived and were starting to move the furniture out and take up the carpet. 'What's going on?' I breathed, as they lugged the *chaise longue* towards the door. Surely Olafson wasn't offloading his foams already? It was one of our biggest problems. Criminals never kept the same pieces of furniture for long. They were always lusting after the latest high-status foam designs. Which meant *we* were always playing catch-up.

On the film, two young men in white disposable overalls chatted about the latest 5D computer game while they attached pipes to a pair of large metal canisters. They fiddled with the nozzles for a moment and then - pretending they were gunning down an enemy attack - aimed them at

the walls. There was a pink blur.   A series of different camera shots kaleidoscoped round the screen as the paths in other parts of the room tried to fill in the lost signal. The sound broke up and was replaced by a static hiss. Within minutes we were looking at a blank screen.

Something *had* gone wrong. Very wrong.

'That's the whole operation gone to waste,' 'Controller Mendez spat. 'Olafson's twigged that we're watching him.'

'Maybe there's another explanation . . . ' But I was just playing for time.

The Controller slapped her hands down on the desk. 'He has all the furniture removed. Then he sprays some kind of *paint* over the walls to block out the paths**.** What other *explanation* is there?' She was shouting now. Specks of spittle peppered the lapels of her spider silk jacket. 'He knows we're on to him. All the foams will have been emptied. He's probably cleaned out his entire network of vases already.'

<div align="center">[SCREEN PAGE BREAK]</div>

As Mendez ranted my brain was working overtime. Something on that film just didn't add up. Those workmen didn't look like gang members; the sort that Olafson would entrust with a big security lockdown. They'd looked – and acted – like regular labourers on an everyday decorating job. And there was something else tugging at my memory. Something had made me think of Lily; her things all bundled into the side pockets in the hive, her pink slipper on the seat, her pink sippy cup.

And suddenly I had the answer. *Pink!*

 'Olafson's married, right?' I asked.

Controller Mendez glared at me but the man in black nodded. 'Yeah. Lucetta Stone. She was a lap dancer in a club the gang took over as one of their first vases when they were just starting out.'

'Kids?' I asked.

The man shook his head. 'We think she may be sick. She's been visiting a medical clinic on Riverside . . .'

'It could be a new vase.' Mendez interrupted.

The man from Internal Investigations shook his head. 'There was no sign she was meeting a contact there. No drop-offs. No deliveries.'

'So she's having a boob job or a laser tuck?' Mendez snapped. 'What's the point of this?'

'Go back to the last frame,' I said, talking to the computer screen. It didn't even flicker. Its voice recognition couldn't deal with my British accent. For once I missed the <u>hive's</u> digital teenager. At least she listened to me.

Mendez sighed, but then turned to the screen. 'Back!' she said, flicking her hand to show how far.

'Stop! That's it!' I cried. 'Look!' The frame froze on a view of the paint canisters. It was blurry – one of the last shots before the <u>paths</u> were blotted out.

Mendez told the computer to zoom in, and there it was. The name of the brand of paint was printed on each canister; *Lullaby Magic.* The colour was labelled too; *Pink Candy Dream.*

I was right!

Lullaby Magic was the best money could buy. The paint had a luminescence that supposedly sparkled in synchrony with the human heartbeat. It was also suffused with soothing tones of vanilla and chamomile for 'unbroken nights of blissful slumber.'

We'd only been able to afford one tiny wall of *Pink Candy Dream* when we decorated Lilly's room.

'I think that Olafson and his wife *are* expecting a delivery at the medical centre.' I said. 'A very special kind.'

Zinnia Mendez and the man in black wore matching blank expressions. I'd have bet my savings that neither had children. 'They haven't rumbled our surveillance,' I explained. 'They're turning that room into a nursery.'

I just hope the Lullaby Magic is more effective on their baby, I thought. 'Blissful slumber' had become something of a family joke between Marla and me. Although at four in the morning neither of us ever found it very funny.

Story 2: Prisons

When they put this pacemaker inside my chest I thought that – well I thought, I thought at least I'd be exempt now; the radiation of the fuel they load into their ships, the magnetic flux of Interspace travel, the half-dead waves still emitted by the hard, petrified star itself. Each of these things would normally shut down, short-circuit and destroy a pacemaker. No, they told me, no, that won't be a problem. And so they opened me up again and the biomedics worked their obscure science and they fixed me up with a protective <u>dawn</u> to act as a guard for my heart, for my pacemaker. This will protect you, they said. This will guard you: the electromagnetic signals will not be able to permeate the <u>dawn</u>; your pacemaker will be safe. They explained it all with words and phrases I could not understand. Well I could not argue with that, and so they shot me off into space along with all the other criminals too low-down to live on Earth any longer.

As a resident of a maximum security prison I was treated as falling *below* the net of society. My crime was so horrific, they tell me, that even I cannot now remember it. They ensured that. Every record of my crime has been wiped, and the military-trained biomedics that served as our Guards have eradicated the event from my mind. Society has licence to ignore me completely; everyone knows that a maximum security inmate belongs to a different order of humanity, that they can experience a sanctioned revulsion at the very thought of us. All of this we know. I know that during my time in maximum security I could not be considered a human, not in the true sense. I accepted this, I still accept this, just as we all accept it.

And this is how they sold us the idea of the mining expedition. The Guards gathered the worst of us in a large well-lit hall, which burned our eyes after countless weeks in the darkness of our cells. The room was filled with charts and maps and projections and a hologram model of a crusty white orb, floating and rotating freely in the centre of the room, about the size of a small transport vehicle. One of the lads pointed to the orb and shouted: 'now that's one pill you biomeds ain't gettin me to swallow.' We all laughed at that, but the Guards did as they always did – they ignored the joke completely, ignored our laughter, until the vacuum of attention became so oppressive we all slowly stopped and looked at each other with something approaching shame.

[SCREEN PAGE BREAK]

'Take a seat' said a woman in a uniform none of us could quite place. And so we all sat around the table that this orb was hovering above. The Guards let us sit in an uncomfortable silence for a long moment while they looked over the files contained on their tablets. One senior biomed emitted a series of short tut-tutting sounds and shook his old head. He was the first to

look at us. 'You're all making us look rather bad' he began. 'Do you know where we are?' He looked around at our faces. 'Do you?'

'In a maximum security prison, sir' said one of the younger men.

'That's right, that's absolutely right' the biomed said over the brim of his glasses. 'And the point of a maximum security prison?'

'To keep us all away from – from the outside, sir, from society, sir.'

'Exactly right; to keep all of you *outside* of society. But you lot are *notorious* aren't you?' We did not know, could not remember anything, and yet we accepted this piece of information as we did all pieces of information in those days. 'Yes, you are all *well-known*. And this poses a problem for us. People cannot forget you as they ought, even with you in here there are those out there who still talk about you, about all of you, in one form or another, and this will not do. We need you *removed* from society… *entirely*.' He closed his eyes for a moment and sighed. 'But never mind that. We have a solution. Yes, do not worry about that. We have found a way to make you all – become forgotten, soon enough. Don't you worry about that.' He smiled and met our gaze. 'We have an exciting proposition for you all.'

A Guard officer pointed to the large spinning white orb. He explained that what we were looking at can only be seen from the most advanced telescopes in society. The kind owned only by the most elite institutions. In fact, he explained, until a very short while ago we had no real idea of what it was at all. Well, now all of that had changed. What we were looking at was death. It was known as a spy, and, it was, as he read from a tablet he held in front of his chest, the unique, hardened remains of a supermassive star – a star which had not become a proud black hole, but which had simply become too large too quickly, had expanded into nothing, and had left behind only a dry core to wither and cool. We were looking at a spy, that's what they called it, and we were going to mine it.

[SCREEN PAGE BREAK]

They showed us all pictures of men and women in their full-body feasts – necessary wear to protect oneself from the radioactive dust thrown up by the mining process. From the singular eye-slit in one man's feast it was just possible to make out the edges of a smile. The miners are all treated very well, we were told. Each of them is free to do as they please on the satellite station orbiting the spy, during their time off. Each and every worker up there, in deep space, mining, is making a singular and worthwhile contribution to society, and are therefore integrated deeply *into* society. This we were told. For many of us this seemed like an impossible dream. We knew we were not accepted as part of the world here in our prison. But

out there, in the frozen pitch of space, we could finally contribute to society and to become a part of it again. Well. I was sceptical.

It was around this time in my life that the biomeds diagnosed my heart condition. They regulated my failing heart with a pacemaker. They cut me open and put this thing inside of me. They told me I wouldn't even notice it was there. That was largely true until they cut me open again and put a dawn over it. The reason for the dawn was to ensure I could make the Interspace expedition to the mining colony, to ensure that my pacemaker did not fail at the first whiff of the spy radiation. Yes, the device was my liberator from the maximum security prison, even if it liberated me for an expedition I did not want to take.

Many of the lads were pretty excited about the whole thing. The day before we left, they gathered us all in to this little white room with medical equipment hanging from the walls. Everyone was whispering their nervous anticipation at the idea of mattering again. And what they told us all then really cemented that sentiment, and everyone felt very good about the whole thing. They told us we would each have to wear a pearl - they wrapped this device around our left biceps, and informed us that they would remain there, under our feasts, for the duration of our expedition to the mining colony. They explained that it would innocuously, without piercing through our skin, unobtrusively, monitor everything it could about us (which, they said, was a great deal more than only a few years ago) via our blood. The information from the pearl would then be sent back through Interspace-waves and arrive safely with someone somewhere back here. We didn't need to worry too much about that, they said, so we didn't. The pearl was quite plain and unassuming and I could barely feel it clinging to my arm, so I really didn't need to give it much attention (in fact it is only now, recounting this part of my tale, that I have given it any thought whatsoever). I could see the faces of the men in the room, their pearl worn proudly around their arms, and I could almost see their tiny, shrivelled, ignored and neglected identities building themselves up again, slowly, knowing that someone, somewhere, was taking an interest in them.

They never actually explained why we had to wear these pearls, but none of us asked so I suppose it was we who were in the wrong. We really should have sensed something when they put these things on us. We should have known, really.

[SCREEN PAGE BREAK]

We took off the next day. They performed a few last-minute checks on all of us. They took us away, one by one, for privacy's sake, into a small medical room and gave us all a once-over. When my turn came, they took me in and placed two electrical prongs over my chest, testing,

they told me, that my <u>dawn</u> was in perfect working order. It was, and I was cleared for take-off.

The shuttle was quite sizeable, and after we had left Earth's orbit, had punctured Conventional space and had slid into Interspace, we were able to move freely around. The expedition would take about a week, they had said. During that time we would all be required to wear our <u>feasts</u>, without the helmet, to protect us from the worst of the high radiation of Interspace. 'This thing is intolerable!' shouted one of the men on the ship. He was complaining about the fact that our <u>feasts</u> were intolerably scratchy. He was complaining that our <u>feasts</u> were intolerably tight in all the wrong places. 'Intolerable, really uncomfortable, I can't wear this thing' he said, as he ineffectually pulled at the suit. We all looked at him then, looked around the sizeable cabin in which we were free to move around, considered the endless free flux of Interspace we were moving through, thought about the mining colony with its meaningfulness and free time. We all looked at him thinking about these things and he saw those thoughts in all of us. That made him quiet down, and no one complained about the really very uncomfortable (and intolerably itchy) suits again.

On the third day I started speaking to one of the other lads on the ship. I pointed to a scar on his head. 'What's that scar on your head all about?' I said.

'Oh this?' he pulled off the glove of his <u>feast</u> and tapped his temple with his finger. 'This is from where they had to implant my, my – my regulator, you know? For my brain. Apparently I am very susceptible to, to, to strokes. I could suffer one any minute, they say. So they had to put this thing in to stop all that.'

'Right,' I said.

'Yeah. It's lucky they diagnosed me in time.'

'Oh yeah?'

'Yeah, they only installed this thing just before we left. One week later and I'd be up here ready to blow a fuse any second.'

'That's lucky,' I said. I felt anxious.

'Well they had to put this other thing in there too.'

I unconsciously grasped my chest.

'Yeah, to stop the radiation or flux or something. They called it – something – they called it – '

'a <u>dawn</u>?'

'That's right. That's what they called it.'

[SCREEN PAGE BREAK]

Well it turns out that we all had one of these things inside of us, for one condition or another. Each of the men (I spoke to them all in turn, quizzed them) had a condition, a device to cure it, and another device, a <u>dawn</u>, with which to protect it. My heart was probably fine, is probably fine, and yet I still have this pacemaker and this other thing inside my chest. And the <u>pearl</u>? I don't know what that blood-monitoring <u>pearl</u> was all about, come to think about it.

I do know how things are now.

When we arrived at the <u>spy</u> we were shown around by someone who kept glancing at his watch. After a few days we saw very few people on the <u>spy's</u> crust. Very few miners. Most of them, we found out, were doing nothing at all. Just wandering around. Eventually we all worked out that none of us actually had to do any mining at all. None of us had to do anything. There were some who would pull on their <u>feast</u> every artificial morning and go out mining, of course. But I do not talk to those people any more. I have not done so for quite some time.

Up here, with my <u>pearl</u> still monitoring everything I do (maybe I should take it off?) I am reminded of something we were told all those years ago in the briefing room with the floating <u>spy</u>. This place can only be seen, they said, from the most advanced telescopes on Earth, owned by only the most elite institutions. Well the only institution which could be considered remotely *elite* is the same one which owns the maximum security prison, which owns everything. Well. It is hard not to feel watched when you think about that.

I can't say I know why I'm here: I don't. Maybe I'm some kind of test subject, maybe I'm not. But I have a funny feeling I'm being watched. Or: I am being watched by some, while others forget. What interest anyone ever had, ever really had, in this <u>spy</u>, or the <u>dawn</u> around my pacemaker, I will never know.

I suppose I know very little, now I come to think of it.

Story 3: Reflections upon a Tribe

The Elder, when he appears, is wearing the ring. He holds his staff aloft and all can see the intricate brass weave on his finger, constricting the small gem which fades through myriad colours. The musicians of the all-male tribal folk band, known as the <u>bruise</u>, stand by the elders of this small tribe, solemn and still, knuckles white wrapped around their rudimentary wind instruments. They patiently await their turn to perform the ceremonial songs of old on their soft-sounding wooden flutes and sheepskin drums.

Stepping forwards against the heavy blue sky, the Elder leads the rather intoxicated villagers (we have all partaken of a potent brew), the dancers, and the musicians of the <u>bruise</u> themselves up to the flatrock overlooking the plain. The sombre group winds slowly up the hillside, and just for a moment one may catch a glimpse of the Elder almost as Abraham himself, leading his only son to his sacrificial death on the mountaintop. I do not know what to expect on the peak, on the flatrock, but the faces of those I have studied now for five long months tell me enough.

The weather is coming in now, and the wind whips up the long dyed gowns of the dancers walking amongst the villagers. Looking up from the base of the winding track, in the deep gloom, the entire tribe appear to comprise a singular dancer; the fluid movements of the tribe's ascent uniformly inimitable. In the daylight, under a clear sky, in the village down here in the lowlands, the street-performers dance an exuberant dance – the <u>fog</u>. The dancers elongate their bodies and sway in fluid motion from side to side, but keep the head still and undisturbed. The <u>fog</u> is one of the many charming qualities, and there are many, of these people. Now, however, under a dark sky heavy with rain, the winding tribe are led by a man disturbed, creeping ever further up the hill in an inverted reflection of the jubilant dance.

[SCREEN PAGE BREAK]

I stay behind awhile on the pretence that I wish to photograph the ascent. This itself is taken with scepticism by the villagers, but the Elder speaks some words I cannot yet grasp and I am tolerated. I assemble and properly arrange my camera, load the film, and take cover underneath its black shroud, holding the trigger ready under my thumb. I hesitate as I watch the familiar <u>fog</u> dance unfold at half speed through the lens. And then for a curious moment, I feel a coldness in my heart. I feel gripped by the impression that I really ought not to take the picture of the <u>fog</u> dancers. I become fearful, anxious. I stare for several seconds, breathless. Somehow my thumb comes down and the loud click of the camera removes me from my reverie.

I uncover myself and turn swiftly away from the procession of villagers, <u>bruise</u>-men and <u>fog</u> dancers, being led towards the summit by the Elder. I shall have the film destroyed, I find

myself thinking. The photograph is not, however, my true reason for the delay in climbing the peak.

Sure that I cannot be seen, I run through the wet grass to my tent. Once there, I begin hastily preparing my scientific instruments for travel. I mean to leave promptly after the ritual on the flatrock; the women will immediately begin the hunting season. The men shall retire to their lodge (the location of which is still a secret to me) for the old, long initiations of the sons of bruise-men into the revered band. It is a strict tribal tradition that when a player retires his closest living male relative should take over his position, which is considered a great honour.

At this point I shall be able to slip away. I must leave. The general feeling appears to have shifted of late. Where once I was warmly welcomed into this small tribe, an aura of anxiety and suspicion has grown malevolently on the horizon. In my dreams these tribespeople, who I must remind myself in my waking hours are little more than savages, become my tormentors. The bruise, who play peaceful folk music on soft wooden flutes and sing of nature in the sun now play an eerie tune and scream pagan incantations in the darkness of my sleeping mind, while the fog becomes not a charming dance, but a tormenting and disturbing display.

My jars clink and test tubes rattle as I swiftly wrap them in napkins and handkerchiefs for the long journey I must retrace through the lowlands. Almost packed, I turn my attention to a small wooden box on the earthen floor. Yes, why not, after all? I think I may permit myself one last experiment.

I open the lid, pull out the concertina of compartments on either side and prop up each of my measuring devices. I unfold a small stretch of paper with my graphs and annotations. Checking the reading on my small barometer and thermometer, I trace my hand along the graph paper until I reach the point where humidity and temperature perfectly intersect. From here I take a third reading. A colour.

[SCREEN PAGE BREAK]

Silently waiting in the middle of my opened wooden box is a velvet handkerchief. I check my readings once again, retrace my fingers along the graph to ensure the certainty of my prediction. Once I am certain, and have readied myself a little for something I can anticipate but not describe, I pull the handkerchief away and see that the cactus, a precious gemstone that is perhaps the most valuable of all the artefacts I have recovered from these tribespeople, is precisely the colour I had expected: a rich burgundy-red.

Back in England, a sample this size of such an extraordinary material shall be highly desired by all major archaeological institutions. The British Geological Survey in particular ought to

offer me a high price. As I take a moment to marvel at the wondrous stone, a fraction of which sits in the Elder's ring, the light diminishes in my tent as the night sky draws in. I turn quickly away from the cactus and check the thermometer. Sure enough, the temperature has dropped a significant fraction of a degree. I consult my chart again, aligning the new data, and produce a colour: deep purple with an orange iris. As though aligning itself to my new reading, the stone changes in less than a second and now appears dark purple with an outside band of orange which bleeds towards the centre. Thus I now feel utterly justified in concluding that the cactus's hitherto mysterious colour transformation is due solely to shifts in atmospheric conditions. Satisfied, I enfold the box back in on itself and pack it away with the rest of my equipment.

It appears to me only now that this presents another reason for my immediate departure: these tribespeople would be much displeased were they to find I had procured their religious artefact in the name of science and civilisation.

With all my instruments packed safely away, I set out up the winding track to the flatrock.

In my mind I begin to write my treatise on these remarkable indigenous folk. I begin to give the lectures which shall accompany the public display of the cactus. 'These remarkable indigenous folk,' I shall say, 'believe that this stone has deep spiritual significance. Yes,' I shall pause, 'during my time with the tribesfolk I was taken in as one of their own. They would talk to me, in their own inimitable rudimentary language, of the significance of the artefact you now see before you' (at which point I shall release the velvet drape to reveal the gemstone on a wonderful neoclassical stand). 'They would tell me tales, old stories delivered verbally (for they are yet to develop a formal writing system) through the generations, of a great monster. The monster, which the old tales name the carton, stalks the lowlands surrounding the village, killing and feeding on the great oxen which graze on the plains' (here I shall gesture to one of my many scientific drawings of these lumbering creatures).

[SCREEN PAGE BREAK]

It is raining heavily now, and my feet slip on the wet mud track. Above me I cannot see the peak beyond the dense foliage. Instead I keep my head down and continue my lecture. 'This monster, ladies and gentlemen, is said to appear as a friend, wearing as it does a wide smile at all times. The true horror of the creature, however, lies herein. The smile of the carton is meaningless. The monster wears the sign of human civility only in mockery, in arrogance. The carton is truly evil, a monster on two legs with a false smile which haunts both the plains

and the dreams of the tribesfolk. Its sickly influence contorts one's nights into a realm of suffering, flooding one's peaceful thoughts with the intoxicating peril hidden behind every facade. The false smile. The horror. One cannot… is unable to… comprehend…' I look down upon the village, cloaked in darkness, and out over the lowlands. The wind howls and the rain beats down, but the twilight is silent. 'This is what the elders say, of course, and nothing more.'

I turn my head to the peak once more. I have reached a turn in the track, and the rain falls down upon me heavier than ever in the clearing. I can see clearly now to the top. The last of the villagers are winding their way onto the flatrock and I know I must hurry. Slipping over wet rocks, I scramble up the track, now covered by dense foliage, now exposed to the biting wind and rain.

I continue to plan my lecture in my mind as I head up the track towards the peak. 'This cactus is of such significance to these people that they believe their gods communicate with them through its changing colours. They believe that because of this, the stone is able to predict the coming of the carton. When the Elder, an old man whom I knew very well, and who was exceedingly fond of me, I must say,' (laugh from the audience at my humorously narcissistic remark) 'holds up his ring on the great flatrock overlooking the plains, it is said that he shall determine whether the monster shall manifest during the coming hunting season… or perhaps whether it is among them already. At this point…' well, I hope to discover what these people do next when I reach the summit and observe the rites for myself.

'Obviously this is quite ludicrous. I have been able to objectively determine that the change in colour of the cactus is causally produced by the most mundane of all phenomena: the climate. Nothing more.' At this point I shall happily detail the research which has led me to this conclusion.

[SCREEN PAGE BREAK]

I near the peak now, and can hear floating notes on the wind, underneath the sound of rainfall. The short high notes of the small woodwinds dance around the deep roll of the long horns; the bruise is already playing their entrancing folk music.

Fog dancers wave their bodies in the rush of wind and music. The catastrophe of weather smashes onto the flatrock. The Elder holds the cactus ring out into the busy gloom with a rigid arm. The whole tribe appear to be performing the fog dance now, subtly shifting their bodies,

trance-like. Booming winds and careful notes battle for dominance. Lightning then thunder begins to break forth from the heavens, the heavy sky at last tearing itself apart. The lecture hall in my mind is swept away and dashed across the lowlands. I have ascended the flatrock to find the village in a frenzy. Each eye is glazed as I try to shout to be heard.

I push my way through the throng, towards the Elder. Around him stands the <u>bruise</u>, in a ring, performing now with maniacal urgency. I shout, but the Elder's attention is transfixed by something high in the sky above. The rain streams across my face. My vision is blurred. I push through the crowd, desperate now, desperate for – for what? I do not know.

'Elder!' I shout, in their language. At this, he starts. At least, I think it is at my call. No, it is something else. The <u>cactus</u> has changed. I look with disbelief. This colour – should it be this colour? I try to recall my charts, my findings, but come up short; my thoughts whip away with the wind. Should it be this colour? The Elder finally looks at me. The whole tribe looks at me. Should it be this colour? 'The <u>carton</u>' he shouts, stepping back. 'The <u>carton</u>!' It takes me a moment to translate. It takes me a little longer to translate the movements of the tribal people. Their gaze. They believe… They believe it is me. And the Elder holds his staff aloft with a fierce look in his eyes.

What happened next I have told not one person. In all my lectures, in all the years since the flatrock, I have never told a soul. I have never told anyone that when the old Elder held up his staff with a glint of madness in his eyes, the crowd forming solid ring all around him and me, the <u>carton</u> embodied, my mind grew cold with fear. I still cannot shake the sensation. I fell to my knees, crying, fearing what would come next, my heart began to beat faster than I ever thought possible. As the <u>bruise's</u> music wrapped around my consciousness and the storm warped my vision, with the dark sky breaking into hellfire above me, the Elder urged me to take hold of his staff. I got to my feet and grabbed hold. At that very moment, a bolt of lightning plunged from the dark sky above and struck the staff. In a flash of blinding white light I was thrown backwards, pain searing through my body. All went quiet. To this day I don't know how I survived it, without a single mark on my body. As I lay there on the flatrock, the storm quickly began to ease, my mind became filled with so many questions – how had the Elder known that the lightning would strike the staff? Surely it could not have been a coincidence. Through the haze of my thoughts, a singular idea was conceived: I was the <u>carton</u> embodied, and the lightning had purged me of the monster and saved the tribe. From *me*. There is no other explanation. There simply is no other explanation.

Story 4: The Island and Elsewhere

We drew into the Island's only dock with the setting sun at our backs. The returning fishing vessels would soon be visible on the horizon. We had followed them the entire way, sitting just out beyond their view. It would have been impossible for them to spot us in our small <u>rug</u>; we had borrowed one of these traditional wooden fishing boats from the dock. We followed the black silhouettes of the large fishing vessels on the edge of the endless sea. Keeping pace was not a problem: <u>rugs</u> are crafted from the wood of the Island's most ancient trees and carved long and narrow, we were gliding through the waters with the speed and agility of a shark. One two-manned <u>rug</u> can skim over calm seas at unimaginable speeds.

 'Are you satisfied Tane?' asked my good friend Maru. 'We have seen it many times now: the large boats do only one thing on the empty, endless sea' he said. 'They fish.'

The girl had been missing now for almost two lunar cycles. I had spoken to nigh every single person living in our <u>rust</u>, and in small island settlements such as our own we rely on each other as if part of one big family. We live in a typical <u>rust</u>: deep inland, almost at the centre of the Island in the only clearing of land, with the dense forest surrounding us all around. The trees give good shelter from the tropical rains, and our huts are huddled tightly together; everyone knows what's going on with everyone else in our close-knit community.

The only person, I soon learned, who had seen anything was an old woman, close to the end of her Island-life. I asked the woman several times if she was sure about what she had told me she had seen from the window of her hut. She insisted.

The old woman had seen the girl, Hana, with a *stranger.*

The world, as we knew it, consisted of only the Island and the endless sea. If you were to set out in a <u>rug</u> and keep a true, straight route across the wide ocean, you would eventually arrive back at the other end of the Island itself, having seen no other landmass, no other person, nothing else in the world but the sea and the Island. We few Islanders were humanity, and we had been blessed with this mound of land and vegetation to live out our long days.

And yet Hana was seen with a stranger.

[SCREEN PAGE BREAK]

I had always been an anomalous presence on the Island. Of all the villagers of the <u>rust</u> throughout our known history, I am the only one never to have been visited by the spirit of one of our ancient ancestors, who reside in the great <u>fee</u>: the vast, flat top of the Island's forest

canopy, where vines and the thick foliage of all the trees are interwoven into one great expanse, stretching on towards the sea.

Because of this I am somewhat distrusted by the others who see me as *different*. So it naturally fell to me to investigate the anomalous, and I have been investigating rare or inexplicable occurrences for most of my life. But never, in all this time, have I ever had to look into something as rare or inexplicable as this.

Hana had been seen with a stranger, someone from *elsewhere*. With no elsewhere in the entire world, finding the missing girl seemed as impossible as her disappearance.

And so I met in secret with my only true friend, Maru. I told him what the old woman had seen and he believed me at once. Together, having become distrustful of the other Islanders, we arranged to follow the fishing fleet daily in one of the small two-manned rugs for as long as it took for us both to be satisfied that they were not dealing with some hidden group of people out there somewhere. Each morning we followed them out beyond the circle of trees protecting the rust, down towards the dock. From here we would watch them embark, wait until the boats were just visible in the distance, discretely board one of the small rugs, and set off in measured pursuit. For many days we continued this pattern, until we had followed the fleet around the entirety of the endless sea, sitting alone in our rug. Only then did Maru ask me if I was satisfied.

'Yes,' I said. 'Yes Maru, I am satisfied.'

'As am I,' he said, gazing out at the growing silhouettes in the distance. He turned to me. 'What will you do now, Tane? Where did this stranger come from, if he did not arrive from some unknown place in the endless sea?'

'I do not know,' I said. 'I shall attend the old ceremonies back in the rust once they begin, for it is a full moon tonight; perhaps nature shall finally open itself to me and I shall gain some valuable insight.'

[SCREEN PAGE BREAK]

The cold damp night blew in from the sea and the forest shivered. As I pushed forwards I came to a familiar opening in the clustered trees; I passed under ancient branches and into the protective warmth of the rust. An orange light held in the air, the fire in the centre of the clearing casting harsh, dancing shadows amongst the leaves of the surrounding trees. The ceremonies had begun.

The drummers controlled the rhythm of the night, which started with a slow but ominous beat. Already many were wearing their ceremonial headdresses, known as <u>cakes</u>, adorned with animal artefacts representing each domain of nature as we knew it; shells for the endless sea, feathers for the sky above the world, and fur for the creatures of the Island. The <u>cake</u> is traditionally worn in religious ceremonies on the island to celebrate our relationship with nature, and in wearing it we conjoin our souls with the flow of the natural world. Through the <u>cake</u> we could hear the Island *speak* to us.

I myself heard nothing, save the wind blowing through the empty shells of my own <u>cake</u>. I saw my fellow Islanders drunk on the hidden fruit of nature, dancing and rejoicing in the purity of life in all things. I thought of Hana and hoped she was alright.

The ceremony lasted long into the night, but I removed my <u>cake</u> early and set off to rest, just as the drums began to herald the real celebrations.

The night was late when I awoke in my resting place under a tall tree, away from the close-knit homes of the <u>rust</u>. With a heavy mind brought on by an unnatural waking I looked around. All was dim about me, save for a faint glow from faraway fires. Sleep came so easily on the Island, and dreams or any disruption of the peace were so rare that I found myself in a mild panic at having awoken before daylight. A little more alert, I sat up and moved my eyes rapidly, searching the thick undergrowth around me. Nothing. At some point I became aware of a slight coldness.

'Hello, Tane' came a voice. I could see, now, a man sitting cross-legged at my feet. I knew the voice and the man who carried it; sitting at my feet was Keola, a man so old that all the Islanders of his time were resting now with the ancestors in the great <u>fee</u> above. He had been cast out from the <u>rust</u> long ago; he had gone insane and the other Islanders soon grew tired of his nonsensical ravings. Now he lived nowhere, roaming the Island and cursing all those unfortunate enough to come across him.

[SCREEN PAGE BREAK]

'My God is forever' he said. 'My God says nothing save what can be spoken, you see?' He pointed to the old ruined <u>cake</u> resting on his head. Keola's <u>cake</u> was unlike any other: adorned on either side with two great shells, curling like horns. He nodded and closed his eyes, 'you see, you see. I see too'. He opened his eyes. 'All is movement. All is resting. These things we know, you see?' He held out his hands, imploring. 'You will see. My God is wide, yes, but my God has eyes only to see what my God may know. My God's being is bottomless. Yes, yes, my God sees all *within* but endlessly it stretches and my God... my God must turn away.

All these things we know,' he looked up, 'but only those who feel the signs may bear it outward. Only those with ears to see and eyes to hear. Do you feel it?' He pushed hard against my chest, then pointed to the canopy above. 'Only those who *can* speak it, do.'

With that I fell into a deep sleep and dreamt of Hana and the stranger. When I awoke the next morning, I knew what to do.

I climbed with the sun up to the canopy's base. There I hesitated, for the great ancestral <u>fee</u> lay above me and I was afraid. With great effort, I pulled myself through the leaves and into another place.

Emerging onto the <u>fee</u>, one sees only vegetation – leaves and vines, all twisted and beautifully interwoven, with a thick layer of mist clinging to the top – endlessly stretching out into the horizon. It is said that were the <u>fee</u> an ocean, not even a fully-paced <u>rug</u> would ever find the end. A cool wind blew across the plane, and the leaves brushed against my arms. In the distance I suddenly heard what sounded like thunder, sustained and growing in intensity. Around me, the branches began to rustle and shake, and a horrific cracking filled the air. It appeared to me that the very substance of the <u>fee</u> was breaking itself apart and grouping into a mass before me. Then, amongst the movement, a shape began to form, looking more and more like a person with each new branch and leaf. I knew what was happening now. Many Islanders had seen it before me. I was being visited by an ancestor.

The ancestor sat a good three heads taller than myself, its head adorned with autumnal leaves arcing back, forming a natural imitation of a <u>cake</u>. 'I knew I would see you this day, my ancestor' I said, bowing my head slightly. 'Old Keola seemed to be trying to tell me-'

'Keola is a fool,' the ancestor spoke, his voice coming from somewhere I could not quite see. 'And, Tane, I am not *your* ancestor. No, your ancestors do not reside with us.' I opened my mouth to speak but the ancestor held up a hand and cut me off. 'I have felt the changing of the winds, we all have, here. There are some of us who feel that the time is come. God is waking. Our work, begun many years ago, may finally be nearing an end. Which is why I come to you now, Tane.'

[SCREEN PAGE BREAK]

My mind began to tilt with the knowledge that my ancestors were not here, were not where I had looked up and spoken in futility all these long years. 'Where else can they be?' I said aloud. 'The Island and the <u>fee</u> and the endless sea, do they not make up the world?'

The ancestor stared deep into my being. 'You are correct, the island and the <u>fee</u> and the endless sea are all that comprise the world. Perhaps… perhaps this is not the entire tale, however. This Island, Tane, hides many secrets. There are those who wish these secrets to be revealed to one such as yourself, such that all may return to order. Change is coming, Tane, the great wave of time washes over us. It only remains to be seen who among us shall be standing once the tide rolls out again.' The ancestor groaned a deep and harrowing groan. 'You must go to the ancestral heart, Tane. All you need to know shall be revealed.' With that, the ancestor let out a great sigh and collapsed into nothingness.

My ascent to the heart of the forest, the ancestral heart, was beset at all times by vines and roots, forming from the undergrowth, pulling at me, back and away. The foliage thickened and spread before me. I could hear the distant sound of drums all around as I neared the centre of the Island. When I finally approached the heart-tree, the most ancient of trees, thick and tall, tremendous roots gripping the earth, I was very nearly defeated. As I neared the heart-tree, however, a chill wind blew from behind.

The old tree creaked and moaned in the wind. I looked up towards the top of the heart-tree, but I could see no further than a few thick branches. When I looked at the base of the tree, many thoughts came into being from the sight – the roots divided into large, hollow cracks before they penetrated the ground. Keola had said that his God's being was bottomless. It didn't occur to me at the time to wonder *which* God he could mean. The same God as the ancestors spoke of waking? The Island itself?

Without thinking, I impulsively strode into one of the dark cracks between the roots.

The track within wound down for what could have been a week, or perhaps an hour or a day. Eventually I arrived at a wide plateau, and stood on its precipice. I looked out over the edge, and saw a new world; a city sprawled out beneath the plateau for as far as I could see, falling away into expansive fields. Above this world hung a veil of mist, and standing there on the edge I could feel the cool wind from the sky. I knew I would find Hana here. The stranger had not come from outside of the Island, they had come from *within* it, and taken her with them.

Appendix C

Table II. Stimulus words and short excerpts of the definitions of their novel meanings used in the two-alternative meaning-to-word training task in the intentional learning condition in Experiment 2 and Experiment 3.

Stimulus Word	Short Excerpts of Novel Meaning Definitions

Story 1: *Pink Candy Dream*

Hive	A new type of compact car manufactured in China.
	A vehicle designed for urban living.
	A car with a small boot but extra storage in side pockets at the front.
Vase	A slang name for a base of criminal operations.
	A command centre used by city crime gangs.
	A headquarters where criminals meet to carry out illegal business deals.
Path	The world's smallest type of surveillance device.
	A gadget that records and feeds back video from a miniscule camera.
	A security instrument that can be moved around via remote control.
Foam	A secure place to store valuables within an item of furniture.
	A safe with a wooden panel disguising the key lock.
	A bespoke handcrafted piece of furniture containing a safe hidden from intruders.

Story 2: *Prisons*

| Dawn | A protective device which is fitted around a pacemaker. |
| | A barrier that protects pacemakers from electromagnetic interference. |

	An implant which acts as a shield against electric and magnetic signals.
Spy	The left-over inner remains of a dead star.
	The unique core left behind by individual stars after they die.
	A star's remnants that can only be viewed using the most powerful telescopes.
Feast	A special protective outfit used for extremely high levels of radiation.
	A full-body suit with only a window to see through.
	A protective suit against radiation that is itchy and uncomfortable to wear.
Pearl	A medical implement that is attached to the body for blood tests.
	An instrument that takes blood measurements without piercing the skin.
	A device that transmits readings from the blood to a receiver in hospital.

Story 3: *Reflections upon a Tribe*

Bruise	A well-regarded all-male group of traditional folk musicians.
	A type of community band which plays traditional folk music.
	A band in which retiring members are replaced by their closest relative.
Fog	A graceful form of dance that is several hundred years old.
	A dance that is typically displayed by street artists.
	A dance performed by extending and swaying the body without moving the head.
Cactus	An especially rare and expensive jewel.
	An unusual type of gemstone often made into jewellery.
	A valuable stone that changes colour with changes in the climate.

Carton	A mythological beast that walks upright on two legs.
	A monster of legend that has a permanent wicked smile.
	A mythical creature that is thought to eat cattle.

Story 4: *The Island and Elsewhere*

Rug	
	An old-style Pacific island fishing boat made of wood.
	A type of small two-man fishing vessel.
	A fishing vessel that can travel quickly on the tranquil sea.

Rust	
	The term for a small type of inland settlement.
	A small village located in a forest glade.
	A village with closely neighbouring houses that is sheltered in a ring of trees.

Fee	
	The flat ceiling of tree-tops over a forest.
	The level roof of the forest canopy where trees and plants are entwined.
	The flat top of the forest canopy where islanders believe their ancestors reside.

Cake	
	A ceremonial headdress worn by people in certain ethnic groups.
	An ornament worn on the head that is adorned with feathers, furs and shells.
	A headdress worn to worship man's place in the natural world.

Appendix D

Table III. Stimulus words and paraphrased versions of the definitions of their novel meanings used in the cued recall of word form test in Experiment 1 and in the meaning-to-word test task in Experiments 2-6. Additional different paraphrased versions of the definitions used in the second meaning-to-word matching test in Experiment 4 only are also listed.

Stimulus Word	Novel Meaning Definitions for Meaning-to-Word Matching Test
Story 1: *Pink Candy Dream*	
Hive	A new type of compact urban car made in China which has anterior side storage space but a small boot.
	A new Chinese-made small car for inner-city living that has reduced boot space but extra storage in side pockets at the front. (Experiment 4 only)
Vase	A slang name for a criminal gang's city headquarters where they meet to carry out illegal deals.
	A colloquial term for a criminal base used by big-city gangs for illegal activities. (Experiment 4 only)
Path	A tiny remote-controlled piece of surveillance equipment that feeds back video from a minute camera.
	The smallest ever surveillance device with video recording and feedback that can be moved by remote control. (Experiment 4 only)
Foam	A uniquely handcrafted item of furniture containing a safe with its lock hidden behind a wooden panel.

A key-locked safe built into a piece of furniture that has been individually handcrafted to be unrecognisable to intruders. (Experiment 4 only)

Story 2: *Prisons*

Dawn

A medical device that is implanted around a pacemaker to shield it against interference from electromagnetic signals.

A biomedical implant fitted around a pacemaker to protect against electromagnetic interference by acting as a barrier. (Experiment 4 only)

Spy

The unique inner remains left behind by dead stars which can only be viewed using the most powerful telescopes.

The distinctive residual core left when a star dies that is only visible through very high-powered telescopes. (Experiment 4 only)

Feast

A protective suit which covers the whole body to guard against harmful radiation but is itchy and uncomfortable.

A full-body suit worn to protect against high levels of radiation but is scratchy and irritating. (Experiment 4 only)

Pearl

A medical instrument that records and transmits readings from the blood without piercing the skin.

A new medical device attached to the body to take blood measurements without using a needle. (Experiment 4 only)

Story 3: *Reflections upon a Tribe*

Bruise

A respected traditional folk band whose male members are replaced by their closest relative when they retire.

An honoured all-male ceremonial band in which retiring musicians are replaced by their closest living relative. (Experiment 4 only)

Fog	An ancient dance in which the dancer, normally a street performer, lengthens and sways their body but holds their head still.
	A dance dating back centuries that is mainly performed by street dancers and involves elongating and swaying the body without moving the head. (Experiment 4 only)
Cactus	A rare precious gemstone used in jewellery that changes colour instantly when the temperature and humidity change.
	A unique and valuable stone that can be used in jewellery and changes colour in seconds depending on air conditions. (Experiment 4 only)
Carton	A beast of folklore that walks upright on two legs, has an evil smile and is believed to eat farm animals.
	A mythical monster that walks on its hind legs and has a mischievous smile, it is said to eat livestock. (Experiment 4 only)

Story 4: *The Island and Elsewhere*

Rug	An old-fashioned two-man wooden fishing vessel that navigates calm seas quickly, it originates from some Pacific islands.
	A traditional Pacific island fishing boat made of wood and operated by two people, it can move fast when the sea is calm. (Experiment 4 only)
Rust	A small settlement in a forest glade with closely clustered houses and trees all around sheltering it.
	A small village in a forest clearing in which houses are close together and sheltered by the surrounding trees. (Experiment 4 only)
Fee	The flat surface of intertwined plants and trees above the forest canopy that is home to islanders' ancestral spirits.

The flat top of the forest canopy, thick with interwoven plants and trees, that islanders believe is the sacred realm of their ancestors. (Experiment 4 only)

| Cake | A headdress traditionally worn by certain peoples in celebration of their relationship to nature, it is made of feathers, furs and shells. |

A ceremonial headdress decorated with feathers, shells, and furs that is worn to celebrate man's place in the natural world. (Experiment 4 only)

Appendix E

Table IV. Target-probe word pairs used in the semantic relatedness judgement task in Experiment 5.

Target Word	Related Probe	Unrelated Probe
Hive	Honey	Bicycle
Vase	Flower	Fox
Path	Trail	Pillow
Foam	Soap	Belt
Dawn	Dusk	Basket
Spy	Mission	Fungus
Feast	Banquet	Skull
Pearl	Jewel	Battery
Bruise	Injury	Address
Fog	Cloud	Blade
Cactus	Plant	Doll
Carton	Juice	Alarm
Rug	Mat	Rocket
Rust	Iron	Comedy
Fee	Payment	Cliff
Cake	Dough	Alien

Appendix F

Table V. Target-probe word pairs used in the semantic relatedness judgement task in Experiment 6.

Target Word	Item Type	Related Probe	Unrelated Probe
Hive	Trained	Honey	Bicycle
Vase	Trained	Flower	Fox
Path	Trained	Trail	Pillow
Foam	Trained	Soap	Belt
Dawn	Trained	Dusk	Basket
Spy	Trained	Mission	Fungus
Feast	Trained	Banquet	Skull
Pearl	Trained	Jewel	Battery
Bruise	Trained	Injury	Address
Fog	Trained	Cloud	Blade
Cactus	Trained	Plant	Doll
Carton	Trained	Juice	Alarm
Rug	Trained	Mat	Rocket
Rust	Trained	Iron	Comedy
Fee	Trained	Payment	Cliff
Cake	Trained	Dough	Alien
Shield	Untrained control	Sword	Baker

Barber	Untrained control	Razor	Basil
Shoe	Untrained control	Sock	Goose
Wool	Untrained control	Cotton	Eagle
Frost	Untrained control	Winter	Golf
Beef	Untrained control	Cow	Blouse
Grain	Untrained control	Rice	Kiss
Torch	Untrained control	Bulb	Elbow

Appendix G

Figure I. Distributions of the raw, log-transformed, and inverse-transformed reaction times in the data for Experiment 5.
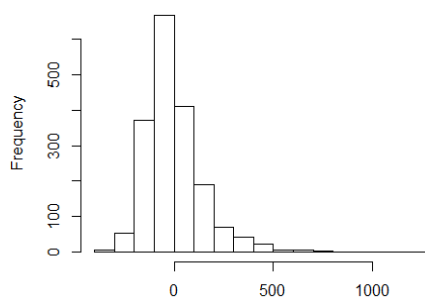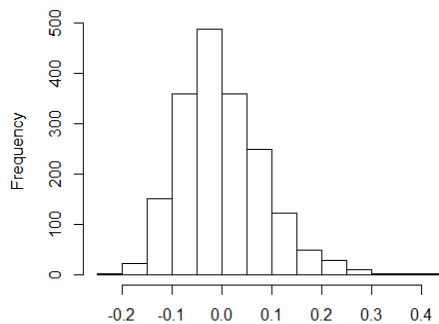
Raw reaction times:
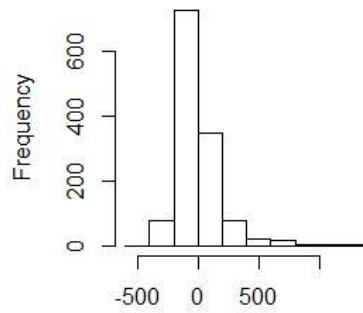
Log-transformed reaction times:
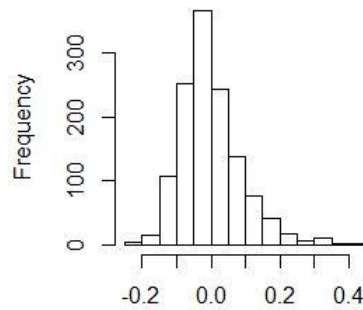
Inverse-transformed reaction times:
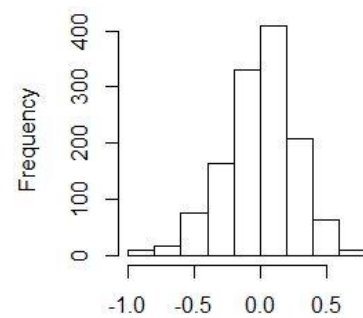
Appendix H

Figure II. Residuals vs. fits scatter plots from the linear mixed effects models for raw, log-transformed, and inverse-transformed reaction times for Experiment 5.

Raw reaction times:
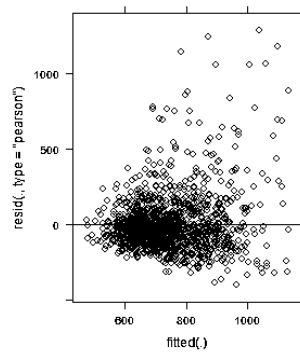
Log-transformed reaction times:
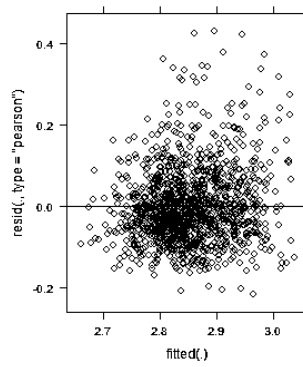
Inverse-transformed reaction times:
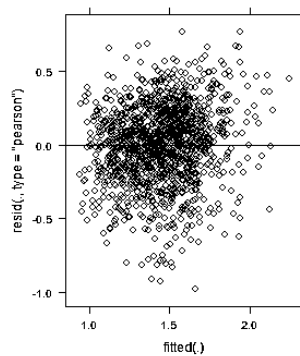
Appendix I

Figure III. Distributions of the raw, log-transformed, and inverse-transformed reaction times in the data for Experiment 6.
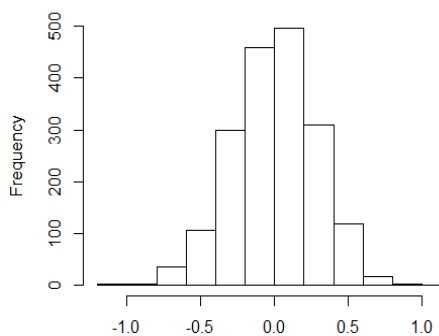
Raw reaction times:
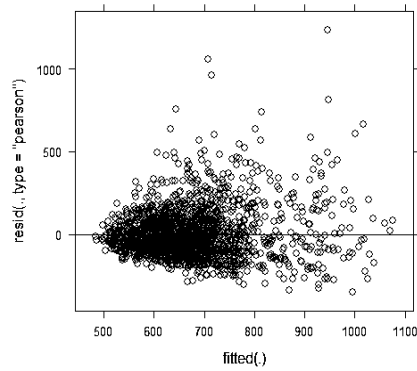

Log-transformed reaction times:

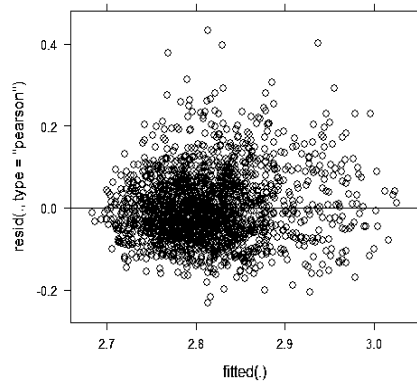Inverse-transformed reaction times:

Appendix J

Figure IV. Residuals vs. fits scatter plots from the linear mixed effects models for raw, log-transformed, and inverse-transformed reaction times for Experiment 6.

Raw reaction times:

Log-transformed reaction times:

Inverse-transformed reaction times: