TALKER-SPECIFICITY AND LEXICAL COMPETITION EFFECTS DURING WORD LEARNING

Helen Brown

Submitted for the award of PhD

University of York

Department of Psychology

December 2011

ABSTRACT

The experiments reported in this thesis examine the time-course of talker-specificity and lexical competition effects during word learning. It is typically assumed that talker-specificity effects depend on access to highly-detailed lexical representations whilst lexical competition effects depend more on abstract, overlapping representations that allow phonologically-similar words to compete during spoken word recognition. By tracking the time-course of these two effects concurrently it was possible to examine the contributions of episodic and abstract representations to recognition and processing of newly-learned words. Results indicated that talker-specific information affected recognition of both novel and existing words immediately after study, and continued to influence recognition of newly-learned words one week later. However, in the delayed test sessions talker information appeared to be less influential during recognition of recently studied existing words and novel words studied in more than one voice. In comparison, lexical competition effects for novel words were absent immediately after study but emerged one day later and remained relatively stable across the course of a week. Together the evidence is most consistent with a hybrid model of lexical representation in which episodic representations are generated rapidly, but robust abstract representations emerge only after a period of sleep-associated offline consolidation. Possible factors contributing to a change in reliance between episodic and abstract representational subsystems include the novelty of an item and the amount of variability in the input during learning. However, talker-specific lexical competition effects were observed in the one week retest, suggesting either that episodic and abstract representations were co-activated during spoken word recognition at this time point, or that perhaps talker information associated with newly-learned words was consolidated in long-term memory alongside phonological information.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

There are many people who have contributed to the work in this thesis. Firstly, I would like to thank my supervisor Gareth Gaskell for his help, support, and advice throughout the course of my PhD, in particular for always encouraging me to rethink, re-evaluate, and try again when things did not go according to plan. The members of my research committee, Graham Hitch and Jelena Mirkovic, and all members of the PRG group (both past and present) must also be thanked for their insightful thoughts and comments, and for helping me to maintain perspective and think about the bigger picture. Gareth Gaskell, Lisa Henderson, Shane Lindsay, and Alex Reid deserve a special thanks for kindly donating their voices (and hours of recording time) to the cause. The experiments, quite literally, would not have been possible without their help! Last but not least, my family and friends must be acknowledged for their continued support, for making sure that I took time off, and for putting up with long absences. Thank you.

The work in this thesis was funded by an ESRC 1+3 studentship.

AUTHOR'S DECLARATION

Data from Experiments 1, 2, and 3 were presented at the Experimental Psychology Meeting, (London, Jan 2010), as well as at the Architectures and Mechanisms for Language Processing Meeting (York, Sept 2010). Experiment 5 was presented as a poster at the Multidisciplinary studies of lexical processing: A workshop for William Marslen-Wilson (Cambridge, July 2011), and as an oral presentation at the Fifth International Conference on Memory (York, Aug 2011). Finally, data from the old/new and male/female categorisation tasks in Experiments 3, 6, and 7, as well as part of the combined analysis reported at the end of Chapter 5, were presented at the 2011 Interspeech Meeting (Florence, Aug 2011), and are included in a short conference proceedings paper, Brown, H. & Gaskell, M.G. (2011). *The time-course of talker-specificity effects for newly-learned pseudowords: Evidence for a hybrid model of lexical representation.* Proceedings of the Interspeech Meeting, Florence, Italy.

CHAPTER 1: MODELS OF LEXICAL REPRESENTATION

One of the key questions that researchers of speech perception must address is how the listener is able to identify words in the face of huge variability in the speech input. Variability results from differences between talkers such as pitch, dialect, intonation, and speech rate, as well as differences in the speech environment such as whether there is background noise. Models of speech perception differ in their explanations of how listeners deal with this variability in the speech input, and whether these extra-linguistic details are stored in memory alongside phonological information. Two key classes of models of speech perception are abstract models and exemplar models.

A key assumption of abstract models of lexical representation is that extra-linguistic information does not affect spoken word recognition. It is assumed that the speech input is reduced to sequences of abstract, ideal phonemes through normalization processes that strip away all perceptually and contextually-specific details (Joos, 1948; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Nusbaum & Magnuson, 1997). Whilst early abstractionist accounts did not explicitly state whether normalization resulted in a loss of extra-linguistic details or whether these details were processed and stored elsewhere (*e.g.*, Joos, 1948), and are perhaps more consistent with hybrid models of lexical representation, later abstract models claimed that normalization results in a loss or discarding of extra-linguistic details (*e.g.*, Pisoni, 1997). An abstractionist account of lexical representation is consistent with the large majority of models of spoken word recognition such as the Cohort Model (Marslen-Wilson, 1987), Distributed Cohort Model (Gaskell & Marslen-Wilson, 1997), TRACE (McClelland & Elman, 1986), and Shortlist (Norris, 1994).

Exemplar models on the other hand assume that all perceptual and contextual details specific to each single occurrence of a word are encoded and stored in memory in detailed episodic traces (Goldinger, 1998; Hintzman, 1986, 1988; Jacoby, 1983a, 1983b; see also Pierrehumbert, 2001, for an exemplar model of speech production). Despite the fact that exemplar models of the lexicon assume that each encounter with a word leaves a unique memory trace, these models are also able to account for perceptual constancy in speech perception. It is assumed that when a spoken word is heard, all stored traces that bear some similarity to the speech input are activated. Subsequently, the representations that are activated most by this

speech input connect to the newly formed episodic representation of the word heard. As a result similar-sounding lexical traces become linked. This is important as it is the statistical clustering of similar traces within the lexicon that allows perceptual constancy to be achieved in exemplar models. Thus, both abstract and exemplar models of lexical representation are able to account for perceptual constancy in speech perception.

Exemplar models however have one key advantage over abstract models since any given speech input provides information not only about the linguistic message, but also about the speaker and the speaking environment (Lachs, McMichael, & Pisoni, 2003). Extra-linguistic information can be used to determine various characteristics of the talker. For example, differences in pitch may provide information about the gender, age, or emotional state of the talker, and in the case of people that you know well it can also serve as a cue to identity. As such, it may often be beneficial to encode and store not only the phonological form of a word, but also information about the talker.

1.1 Examining the claims of abstract and exemplar models

A number of researchers have investigated whether information about extra-linguistic details such as talker identity, speech rate, and amplitude (collectively referred to as *indexical* information) are encoded and stored within the lexicon. There are multiple sources of evidence supporting the claim that indexical information is retained in memory, consistent with the predictions of exemplar models. Firstly, it has been demonstrated that identification of spoken words benefits from familiarity with the talker (Nygaard, Sommers, & Pisoni, 1994), indicating that information about specific talker attributes are retained within lexical memory and aid later processing of lexical items produced by that talker. Secondly, familiar voices are easily recognizable even when the input signal is reduced to sine-wave speech (Remez, Fellowes, & Rubin, 1997) demonstrating that listeners are able to use their knowledge about the vocal characteristics of different talkers, acquired during previous encounters with those talkers, to identify talker-specific differences in the phonetic properties of the incoming speech signals.

More importantly, indexical information appears to be stored in memory with links to specific lexical items. Numerous studies have shown that changes in indexical information between study and test can affect the recognition and

processing of existing words when young adults are re-exposed to those words a few seconds or minutes later (Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; Goh, 2005; Goldinger, 1996; Goldinger, Kleider, & Shelley, 1999; McLennan & Luce, 2005; Schacter & Church, 1992; Sheffert, 1998), consistent with predictions of exemplar models. Similar effects have also been observed in infants (Houston & Jusczyk, 2003) and older adults (*e.g.*, Sommers, 1999), as well as using newly-learned words (Creel, Aslin, & Tanenhaus, 2008; Creel & Tumlin, 2009, 2011). Moreover, talker-specific information appears to be retained, to some degree, in long-term memory, and has been shown to affect identification of single words when heard in a background of white noise up to one week after initially encountering a word (Goldinger, 1996), suggesting not only that indexical information is encoded and stored during the initial encounter with a lexical item, but that this information is retained in memory and is linked to specific lexical items for a considerable period of time after initial encoding.

One possible explanation of talker-specificity effects (TSEs) is that different connotations of a word may be activated depending on whether an item is heard in a male or a female voice (*voice connotation hypothesis*; Geiselman & Crawley, 1983). However, there is evidence arguing against this hypothesis, instead suggesting that it is the specific acoustic properties of each talker that are retained in memory and affect later recognition of recently-encountered words. Firstly, Palmeri, Goldinger, and Pisoni (1993) observed TSEs for within-gender changes as well as for between-gender changes at test (see also Sheffert & Fowler, 1995). If TSEs had been driven by different gender connotations for each word then within-gender changes should not have affected performance. Secondly, Goldinger (1996) used multidimensional scaling to demonstrate that it is the perceptual distance between voices that determines performance in different-talker test trials, not the gender of the talker. In other words, voices that were perceptually more distinct produced stronger specificity effects.

Further evidence against a voice connotation hypothesis and in favour of an exemplar-based explanation comes from the finding that allophonic information (*e.g.*, voice onset time, vowel duration, *etc*.) has also been found to affect the processing of existing words (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; McMurray, Spivey, Aslin, Tanenhaus, & Subik, 2008; McMurray, Tanenhaus, & Aslin, 2002; Shatzman & McQueen, 2006) as well as recognition of these words

when presented again several minutes later (Ju & Luce, 2006). Perceptual learning effects, in which repeated exposure to words containing atypical phonemes (*e.g.*, the phoneme from the mid-point of the /s/-/ʃ/ continuum) results in shifts in phoneme boundaries (Norris, McQueen, & Cutler, 2003; Scharenborg, Mitterer, & McQueen, 2011), appear to be robust for a 12 hour period after initially encountering the experimental items (Eisner & McQueen, 2006) indicating that allophonic information, like indexical information, is retained in long-term memory. In fact, Kraljic and Samuel (2005) found that perceptual learning effects were actually enhanced after a period of 25 minutes. The authors suggest that this finding demonstrates stabilization of allophonic information over time and that allophonic information may become part of the stored phonemic representation.

A number of research areas outside the domain of speech perception also provide evidence that episodic details are retained in memory and can affect later recognition and processing of previously encountered items. Firstly, surface form details appear to be retained in memory when written words are encountered, and can affect subsequent processing and recognition of those written words several minutes later (Brown & Carr, 1993; Goldinger, Azuma, Kleider, & Holmes, 2003; Graf & Ryan, 1990; Hintzman, Block, & Inskeep, 1972; Jacoby & Hayman, 1987; Kirsner, 1973; Kolers & Ostry, 1974; Roediger & Blaxton, 1987; Tenpenny, 1995). Font information, like indexical and allophonic information, also appears to be retained in long-term memory; for example, Craik and Gemar (as cited in Craik, 1991) found that participants read passages of text more quickly when the passage had been presented in the same typology one week earlier.

In addition to these demonstrations of surface-form specificity effects participants are also often able to remember the presentation modality of words (Kirsner, 1974; Light, Stansbury, Rubin, & Linde, 1973), the exact wording of sentences (Begg, 1971; Keenan, MacWhinney, & Mayhew, 1977), and even the spatial location of information in a text (Lovelace & Southall, 1983; Rothkopf, 1971). Moreover, a number of researchers have demonstrated retention of episodic details in non-linguistic domains, including memory for musical pitch and tempo (Halpern, 1989; Levitin & Cook, 1996), as well as pictures and visual scenes (Konkle, Brady, Alvarez, & Oliva, 2010; Snodgrass, Hirshman, & Fan, 1996). These latter findings suggest that indexical and allophonic specificity effects may be

accounted for by broader models of memory that assume episodic representation in all domains (Eich, 1982; Gillund & Shiffrin, 1984; Nosofsky, 1991; Tulving, 1983; Underwood, 1969).

However, whilst the findings outlined above clearly support exemplar models of the lexicon, it appears that not all indexical information is retained within the lexicon; some surface variables such as amplitude do not appear to affect memory for spoken words, whilst other surface variables such as speech rate and talker identity do (Bradlow et al., 1999; Church & Schacter, 1994; Sommers, Nygaard, & Pisoni, 1994). One possible explanation for this dissociation is that changes in talker or speech rate can affect the phonetic properties of a word such as voice onset time or vowel quality and duration, whereas amplitude changes do not (Remez et al., 1997), suggesting that perhaps only functionally relevant indexical information is retained within the lexicon. In support of this suggestion Kraljic, Samuel, and Brennan (2008) demonstrated that odd, mid-point pronunciations of the phonemes /s/ and /f/ affected later phoneme categorization only when the pronunciations were deemed to be characteristic of the talker. When participants heard the odd phonemes alongside a video in which the talker transiently put a pen in her mouth whilst pronouncing the words, later phoneme categorization was not affected. It is also important to note that not all studies have shown robust font-specificity effects; some studies have failed to find font-specificity effects at all (*e.g.*, Gibson, Brooks, Friedman, & Yesavage, 1993), whilst others have shown only small and inconsistent effects (*e.g.*, Jacoby & Hayman, 1987).

However, it is possible that some of the null specificity effects may simply have arisen as a function of the stimuli and tasks used in the experiments, not as a result of lexical abstraction; it may be that extra-linguistic information was stored by participants but that the experimental measures used were not sensitive enough to tap into this information during retrieval, or that access to this extra-linguistic information was not required in order for the task to be completed. In support of this suggestion, when Bradlow et al. (1999) required participants to make explicit same/different judgments for words classified as old previously studied words, participants were able to accurately identify changes in amplitude, just as they were able to identify changes in talker-identity and speech rate. Thus, it is possible that the retrieval of indexical information may in fact be a deliberate or intentional strategy that is dependent on explicit recall strategies at the decision stage of spoken word

processing (Creel et al., 2008). There are three main sources of evidence that support this idea. Firstly, Magnuson and Nusbaum (2007) showed that only participants expecting to hear two different talkers showed a processing cost in a speeded target monitoring task. Participants who were told that they would only hear one talker did not show processing costs when in fact two talkers were heard. Secondly, Kouider and Dupoux (2005) demonstrated that when time-compressed speech was used in subliminal priming, TSEs were not observed, further supporting the idea that indexical-specificity effects occur at a conscious information processing stage rather than occurring implicitly. Finally, Pilotti, Bergman, Gallo, Sommers, and Roediger (2000) demonstrated that when only one voice was used during test indexical-specificity effects were not found, presumably because when all test items are spoken by the same talker voice information is rendered less informative as a cue to spoken-word retrieval, and thus participants place more emphasis on the abstract phonological information contained within the speech signal (Goh, 2005). Therefore, it appears that whilst all indexical information may be stored, different types of indexical information may be used to different extents depending on the demands of the task.

Moreover, whilst it seems clear that extra-linguistic information such as talker identity, speech rate, and font, is able to influence recognition of recently encountered items (Bradlow et al., 1999; Craik & Kirsner, 1974; Creel et al., 2008; Creel & Tumlin, 2009, 2011; Goh, 2005; Goldinger, 1996; Goldinger et al., 1999; McLennan & Luce, 2005; Schacter & Church, 1992; Sheffert, 1998) one limitation of this literature is that few studies have examined indexical-specificity effects over time for existing words and those that have provide mixed evidence as to whether these extra-linguistic details are truly retained in long-term memory. Evidence supporting maintenance of episodic details in long-term memory comes from a study by Goldinger (1996) showing significant TSEs in an identification-in-noise task one week after initially studying a set of existing words, as well as a study by Ernestus (2009) demonstrated that information about unreduced vowels in newly-encountered past participles could affect recognition of those items in a lexical decision task one week later. These findings suggest that representations in long-term memory may be episodic in nature. On the other hand Goldinger failed to demonstrate sustained TSEs one week after study in a more explicit old/new categorisation task despite

using the same stimuli as used in the identification-in-noise task. This latter finding suggests that representations in long-term memory may be abstract in nature.

Further evidence supporting abstract representation in long-term memory comes from studies by McQueen and colleagues examining lexically-driven perceptual learning effects (Cutler, Eisner, McQueen, & Norris, 2006; McQueen, Cutler, & Norris, 2006). In these experiments participants were exposed to the ambiguous phoneme /?/ midway between the phonemes /f/ and /s/ in the context of either /f/-final or /s/-final words. At test a cross-modal priming task was used in which auditory primes containing the ambiguous phoneme /?/ were followed by visual targets that could be either /f/-final or /s/-final words. Exposure to /?/ in the context of /f/-final words during study biased participants to interpret /?/ as /f/ in the cross-modal priming task, and resulted in greater priming for /f/-final compared to /s/-final items, and vice-versa in the /s/-final exposure condition. Importantly, none of the items used during the cross-modal priming task had been heard during exposure. Similar generalisation of perceptual learning to untrained items has also been demonstrated using noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005) and accented speech containing a vowel shift (Maye, Aslin, & Tanenhaus, 2008). Generalisation of perceptual learning effects to untrained items argues against an episodic basis for perceptual learning effects, instead suggesting that perceptual retuning occurs at an abstract pre-lexical level (Cutler, 2008; McQueen et al., 2006).

1.2 Hybrid models of lexical representation

A number of researchers have proposed that it is necessary to combine aspects of both abstract and exemplar models of lexical representation in order to fully explain the mixture of episodic and abstract effects outlined above (Bowers, 2000; Feustel, Shiffrin, & Salasoo, 1983; Graf & Ryan, 1990; Jacoby & Hayman, 1987). These hybrid models of lexical representation typically assume that new lexical representations are episodic in nature, with multiple episodes combining into more abstract units over time, given multiple exposures to a specific word (Feustel et al., 1983; Goldinger, 2007). Importantly, hybrid models do not assume that episodic representations are lost once more abstract representations have been formed. Rather, episodic and abstract representations are thought to co-exist in memory. Distributed models of memory (*e.g.*, McClelland & Rumelhart, 1985) offer a similar

explanation, with traces of specific, individual experiences represented as unique patterns of activation across a number of nodes in a connectionist network and abstraction and generalisation emerging from the superposition of similar memory traces. The advantage of hybrid (and distributed) models is that abstract representations provide the stability and phonetic constancy that is key to abstract models, whilst highly-detailed episodic representations allow indexical information to be retained, accounting for the indexical-specificity effects that have been observed in the literature.

Evidence supporting the suggestion that there may be both abstract and episodic representations within the lexicon comes from work by Pilotti et al. (2000) who noted that whilst changes in indexical features from study to test reduced priming they did not completely eliminate it, indicating that both episodic and abstract representations play a role in spoken word recognition. In addition, whilst Orfanidou, Davis, Ford and Marslen-Wilson (2011) failed to find differences in repetition priming for same- and different-voice repetitions, suggesting that voice information was not accessed in this task, they did find repetition priming for both existing words and for pseudowords. The authors concluded that aspects of both abstract and episodic theories need to be considered in order to fully explain their results, since abstract models typically predict that repetition priming can occur only if there is an existing lexical representation for that item and thus should not predict repetition priming for pseudowords (see Feustel et al., 1983, for similar findings with written words). However, given that Orfanidou et al. used 'word-like' pseudowords (*e.g., drow, thod*) that were composed of portions of familiar words an alternative explanation may be that the observed pseudoword priming effects in this experiment were driven by abstract sublexical representations (e.g., representations of morphemes or syllables; Bowers, 2000; Dorfman 1994), rather than newly-generated lexical representations, as would be predicted by episodic models.

Nonetheless, there is evidence that episodic and abstract effects may be mediated by different hemispheres in the brain, with episodic effects depending more on the right hemisphere (RH), and abstract effects depending more on the left hemisphere (LH; *e.g.,* Marsolek, 1999). For example, Gonzalez and McLennan (2007) examined hemispheric differences in TSEs for existing words and found that long-term repetition priming was greater for target words presented in the same voice as the prime compared to targets presented in a different voice only when the

target word was presented to the left ear and thus via contralateral neural projections to the RH. This RH processing advantage for episodic details does not appear to be specific to talker information. Gonzalez and McLennan (2009) later demonstrated that long-term repetition priming in the RH, but not the LH, was greater for environmental sounds (*e.g.*, thunder, police siren, etc.) when the same exemplar of a sound was presented in both the prime and target blocks (*i.e.*, the same police siren), compared to when two different exemplars of the same sound were presented in each block (*i.e.*, two different police sirens). Further evidence supporting the suggestion that abstract and episodic information are stored separately within the brain comes from neuropsychological research into phonagnosia, a neurological disorder that selectively impairs voice recognition whilst sparing speech perception, suggesting that voice and word-specific information are stored independently, and that the two can be disrupted separately (Vanlancker, Cummings, Kreiman, & Dobkin, 1988).

Notably, studies in the visual domain have demonstrated a similar RH advantage for the processing of same-exemplar items compared to different-exemplar items (Burgund & Marsolek, 1997; Marsolek, 1999, 2004; Marsolek & Burgund, 2008; Marsolek, Kosslyn, & Squire, 1992; Marsolek, Squire, Kosslyn, & Lulenski, 1994). The similar pattern of hemispheric asymmetry across modalities suggests that hemispheric asymmetries in processing may reflect a more general property of perceptual processing. Nonetheless, it is important to note that Marsolek and colleagues do not assume a complete dissociation between processing in RH and LH. Rather, they assume that both types of processing can operate in parallel in both hemispheres, but that different types of processing operate more, or less, effectively in each hemisphere (Marsolek, Schacter, & Nicholas, 1996).

Given the convergence of evidence towards a hybrid model of lexical representation, one question that remains is to determine under which circumstances each type of representation is more or less likely to be involved in memory recall and recognition processes (McLennan, 2007). Possible variables mediating which system is used include the type of task used, the time-course of processing, the typicality of the stimuli, and the novelty of an item (Marsolek, 2004).

With regards to a *time-course hypothesis*[1], Luce, McLennan, and Charles-Luce's (2003; see also McLennan & Luce, 2005) have proposed that indexical information only affects spoken word recognition when processing is slow, as is the case in tasks that require more explicit responses. Alternatively the difficulty of the task, and thus the speed of processing, may be determined by the nature of the words themselves (Luce et al., 2003; Mattys & Liss, 2008; McLennan & Luce, 2005). Typically, when short, highly-frequent words are used episodic effects are not seen (Luce & Lyons, 1998), presumably because these words are processed quickly and with little effort. When longer, lower-frequency words are used episodic effects become apparent (Luce & Lyons, 1999). Similarly, words with low phonotactic probability, but not words with high phonotactic probability, are affected to a greater extent by changes in talker and speech rate since the former are typically processed much slower than the latter (McLennan & Luce, 2005). Indeed, it has been suggested that episodic traces show greater involvement in word recognition when processing atypical or unusual word forms. Evidence supporting this suggestion comes from a study by Brown and Carr (1993) who found abstract priming when typical forms (typed words) were presented, but font-specific priming when atypical forms (handwritten words) were processed (see also Graf & Ryan, 1990). Likewise, Nygaard, Burt, and Queen (2000) showed that not all forms of surface variability affected recognition of spoken words to the same extent, rather items that were judged as more typical in speech rate, amplitude, and vocal effort showed smaller surface-form repetition benefits. Taken together, these findings suggest that abstract codes dominate spoken word recognition during rapid processing, and that indexical details are only integrated with the retrieved abstract representation when a stimulus is either phonetically or lexically ambiguous, where further information is required to disambiguate the item, and thus where the overall processing time is longer (McLennan, Luce, & Charles-Luce, 2003). It is thought that abstract representations dominate during rapid processing because they code the most frequent features, aggregated across all instances of a word, and thus resonate more strongly and quickly than episodic codes (Goldinger, Pisoni, & Logan, 1991).

[1] Note, *'time-course'* is used interchangeably throughout the thesis to refer to two different aspects of processing. For the most part 'time-course' refers to changes in talker-specificity and lexical competition effects over the course of a week. However, 'time-course' is also used to refer to the time-scale with which episodic details are integrated into retrieved abstract representations during online processing. This latter use of 'time-course' is necessary in order to maintain consistency with Luce et al.s (2003) terminology (i.e., their *time-course hypothesis*).

One interesting point to note here that it is increased processing time, not increased processing effort that is the key to demonstrating episodic effects (Mattys & Liss, 2008). In support of this claim, McLennan and Luce (2005) have demonstrated that speech rate affected the processing of bi-syllabic words and non-words in a delayed- but not an immediate-shadowing task. In this study the same words were used in both tasks and thus required the same amount of processing effort, ruling out the possibility that indexical-specificity effects are dependent on processing effort as compared to processing time. Taken together, these findings suggest that abstract and episodic representations may co-exist, with episodic representations requiring more time to be retrieved and integrated with the abstract phonemic representations.

An alternative possibility is that episodic traces contribute more to processing when items do not have pre-established representations (Feustel et al., 1983; Orfanidou et al., 2011), or are loosely specified in long-term lexical memory. Evidence supporting this suggesting comes from a study by Bowers et al. (1996) who found same-case priming for printed pseudowords, but not for existing words. The authors suggested that the episodic system contributes to processing only when normal access to orthographic representations is prohibited, or when there is no existing lexical representation. Similarly Jacoby and Hayman (1987) argue that many studies demonstrating surface-form specificity effects have used tasks or materials for which people lack expertise. Thus, specificity effects may emerge only because the abstract representations that would free one from reliance on visual details have not yet been acquired.

The experiments reported in this thesis attempt to disentangle the contributions of episodic and abstract representations in lexical memory by examining TSEs for novel words at different time points. One of the main advantages of using novel words rather than existing words is that it is possible to control the number of prior exposures to each item, and the amount of variability in these exposures, allowing more stringent tests of the predictions of abstract and exemplar models of lexical representation than is possible with existing items that not only have pre-established lexical representations, but have also been encountered many times prior to an experiment. Moreover, it is possible to track changes in the contribution of abstract and episodic representations during recognition of novel words as they gradually become established in long-term memory. Chapter 2 describes findings from

previous studies of word learning and outlines a framework that has not only been used to account for these previous findings, but which may also provide a framework within which aspects of episodic and abstract representation may be combined. The remaining chapters outline a series of experiments designed to test the predictions of this framework.

CHAPTER 2: WORD LEARNING, MEMORY CONSOLIDATION, AND COMPLEMENTARY LEARNING SYSTEMS

Word learning is one of the key components of language acquisition. Typically children acquire their first words between the ages of 10 and 14 months (Horst, McMurray, & Samuelson, 2005). By two years of age a child will know approximately 300 words, with this number increasing to over 14,000 by a child's sixth birthday (Carey, 1978), and reaching approximately 30,000 by adulthood (Altmann, 1997). Whilst it is commonly assumed that most word learning occurs during childhood and adolescence, it is clear that adults continue to acquire new lexical forms throughout their lifetime, particularly scientific or technological terms that are associate with one's occupation (*e.g., polysomnography*), or new terms that are introduced as a result of technological advances (*e.g., skype, blogging*). Adult learners may also acquire new lexical items as a result of learning a second language. Thus, it is important to understand the processes involved in word learning in adulthood, as well as those involved in word learning during childhood.

2.1 The time-course of word-learning

In both children and adults it is commonly assumed that learning a new word is a relatively rapid process that is complete almost immediately. This assumption follows from Carey's (1978) theory of fast mapping, which proposed that children were able to infer the meaning of new words after only minimal exposure. Rapid acquisition of novel object names has been observed in both children and adults (Markson & Bloom, 1997). Yet Carey did not assume that fast-mapping was the end point of word learning. She claimed that fast-mapping must allow the child to create an initial representation of a word that contained sufficient information to allow this representation to be maintained within the lexicon until a more stable and complete representation could be developed through further experience with that word (Carey, 1978; see also McGregor, Friedman, Reilly, & Newman, 2002). In the developmental literature, this process of establishing more robust representations has been referred to as *slow mapping* (Capone & McGregor, 2005; Carey, 1978; Horst, McMurray, & Samuelson, 2006).

Recent research in adults also suggests that word learning is not complete immediately after studying a novel item, and has shown that different aspects of word learning emerge at different time points. Leach and Samuel (2007) have drawn a distinction between two different aspects of word learning, *lexical configuration* and *lexical engagement*. Lexical configuration involves learning a words form (phonological and orthographic), its meaning, and which syntactic category the word belongs to. Lexical engagement on the other hand refers to the ability of a novel word to interact with and affect the processing of existing lexical items. Whilst it is useful to consider the differences between these two aspects of word learning it appears that some aspects of both configuration and engagement emerge immediately after study, whilst others emerge after a delay, suggesting that both lexical configuration and lexical engagement of novel words change over time.

Immediately after initial exposure to new words there is clear evidence of robust form-based learning of both novel phonological forms (Davis, Di Betta, Macdonald, & Gaskell, 2009; Dumay & Gaskell, 2007; Dumay, Gaskell, & Feng, 2004; Gaskell & Dumay, 2003; Snoeren, Gaskell, & Di Betta, 2009; Tamminen & Gaskell, 2008) and novel orthographic forms (Bowers, Davis, & Hanley, 2005; Clay, Bowers, Davis, & Hanley, 2007). This learning is essential since some learning must occur immediately in order for further learning to aid the establishment of a more stable representation in long-term memory. As a result of this rapid form-based learning novel words are able to exert 'top-down' influences on perceptual learning of phoneme boundaries (Leach & Samuel, 2007), are able to bias phoneme categorization in a Ganong task (Sedin, 2006), and can influence phoneme judgments in the context of compensation for assimilation (Snoeren et al., 2009), all immediately after a novel word has been learned. Semantic information can also be extracted from single encounters with novel nonwords in cases where that nonword is presented in a highly constraining sentence context (Borovsky, Elman, & Kutas, 2010; Borovsky, Kutas, & Elman, 2010).

However, it is only after a period of sleep-associated offline consolidation that novel words begin to engage in lexical competition with similar sounding words (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen, 2010; Tamminen & Gaskell, 2008; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010). Engagement in lexical competition is thought to be an indicator that the novel word has established a representation in long-term memory,

with this novel representation being integrated with existing lexical representations (Gaskell & Dumay, 2003). According to the Cohort model of spoken word recognition (Marslen-Wilson & Zwitserlood, 1989) lexical competition occurs between phonologically similar words up to the point at which only one word in the lexicon matches the speech input. This point is known as the uniqueness point. Previous experiments have taught adults novel nonwords (*e.g.*, *cathedruke*) that differed from their existing basewords (*e.g.*, *cathedral*) only after the normal uniqueness point, thus shifting the uniqueness point of the baseword towards its offset (Davis et al., 2009; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008; Tamminen et al., 2010). Once the novel nonword has become integrated into the lexicon there should be a greater amount of lexical competition when the baseword is heard, thus slowing processing of that existing item.

Initial studies showed that the same novel nonwords that rapidly generated novel phonological representations did not engage in lexical competition with their basewords immediately after study, but did so the following day (Dumay et al., 2004; Gaskell & Dumay, 2003). One limitation of these initial experiments is that they were unable to differentiate between the effects of increased time between study and test, and effects of sleep-associated offline consolidation. Dumay and Gaskell (2007) carried out a study that enabled differentiation of these two possibilities. Participants studied novel nonword at either 8am or 8pm, and completed tests of lexical competition both immediately after study as well as 12 hours later. Critically, lexical competition effects emerged only for those participants that had slept during the 12-hour interval between study and the delayed test. This experiment provides evidence that sleep-associated offline consolidation is important for the emergence of lexical competition between newly-learned words and phonologically-similar existing words. However, recent research suggests that sleep may not always be necessary for the integration of existing and novel words in long-term memory. Lindsay and Gaskell (submitted) found engagement in lexical competition prior to sleep when participants were trained and tested at multiple time-points across a single day. The authors suggest that spaced learning and testing may, like offline-replay of recently-learned information during sleep, allow novel information to be integrated into long-term memory. Nevertheless, the largest increase in lexical competition effects still emerged when participants were tested the following day, after a period of sleep, suggesting that sleep may be the optimal state during which

integration of existing and novel information occurs. Importantly, the increase in lexical competition observed after learning a phonologically-similar nonword is a robust, long-lasting effect that can still be observed several months after initially encountering a novel nonwords (Tamminen & Gaskell, 2008).

A further example of novel words demonstrating long-lasting competition effects comes from a study by McKay, Davis, Savage, and Castles (2008) in which participants were trained on the phonology, orthography and semantics of novel nonwords, and were tested on their ability to recognize and define these newly-learned words, as well as their ability to read aloud the newly trained words and similarly-spelled existing words. Critically half of the novel items had consistent spelling-to-sound mappings, whilst the other half had inconsistent spelling-to-sound mappings. When tested between 6 and 12 months after initially learning the words it was found that novel items with inconsistent spelling-to-sound mappings still affected pronunciation of existing similarly-spelled words suggesting that novel words engaged in competitive processes even after an extended delay between study and test.

One interesting point to note is that the lexical competition effects, as measured using lexical decision, do not appear to emerge more quickly or to be of a larger magnitude when semantic information is supplied during exposure to the novel words (Dumay et al., 2004, but see Leach & Samuel, 2007 for effects of semantic training on the emergence of perceptual learning effects). Nevertheless, the integration of novel words with semantically-similar existing items in long-term memory appears to undergo a period of consolidation; for example Clay et al. (2007) found that novel words did not slow down processing of semantically related items in a picture-word interference task until one week after participants had initially learned the novel words. Tamminen (2010) also demonstrated that consolidation of semantic information emerges slowly, with semantic priming using visible nonword primes observable one day after initial exposure to a novel word, but masked semantic priming effects emerging only one week later.

There are a number of findings suggesting that consolidation processes may also play a role in the stabilization and enhancement of novel form-based representations in addition to being important for the emergence of lexical competition. Firstly, free recall accuracy for novel words has been shown to improve after a period of sleep-associated offline consolidation in adults (Dumay & Gaskell,

2007; Tamminen et al., 2010), as has the speed with which novel words are recognized (Snoeren et al., 2009) and repeated (Davis et al., 2009; Gagnepain, Henson, & Davis, 2010). Similarly, cued recall of new words has been found to improve overnight in 7 to 13 year old children (Brown, Weighall, Henderson, & Gaskell, in press; Henderson, Weighall, Brown, & Gaskell, submitted-a, submitted-b), whilst recognition of novel words has been found to improve one week after initial training in 3 to 6 year old children (Storkel, 2001), even with no additional exposure to the novel words after the initial study phase in each of these experiments. Further studies have shown improvements in recognition of novel words 4 (Frazier Norbury, Griffiths, & Nation, 2010) and 24 weeks after initial exposure in 6 to 7 year olds (Dockrell, Braisby, & Best, 2007).

2.2 A complementary learning systems account of word learning

The temporal dissociation between the different aspects of word learning described above has been interpreted within dual-system models of memory and learning, such as the *complementary learning systems (CLS)* framework (Davis & Gaskell, 2009; McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003; O'Reilly, 2001; O'Reilly & Norman, 2002; O'Reilly & Rudy, 2000). The CLS framework proposes that newly encoded memory representations are initially mediated by the hippocampal network, but that during sleep-associated offline periods these representations are replayed, resulting in the strengthening of neocortical representations and integration of new memory representations with information already held in long-term memory (see Marr, 1970; 1971, for a precursor to the CLS framework). It is assumed that the hippocampal system forms pattern-separated, non-overlapping representations, allowing new information to be stored rapidly in a way that does not interfere with existing information. In contrast, within the neocortical network, information is encoded more slowly, allowing gradual integration of new and existing information through the formation of overlapping, distributed representations that are sensitive to shared structure between different representations (Davis & Gaskell, 2009; Norman & O'Reilly, 2003; O'Reilly, 2001; O'Reilly & Rudy, 2000). One key point to note is that the CLS framework does not necessarily assume that memory relies on either the hippocampus or the neocortex alone; the two learning systems are best conceptualised as separate but interlinked and highly interactive systems.

In support of the CLS framework a recent fMRI study demonstrated different neural responses to novel nonwords learnt one day prior to the fMRI session (consolidated nonwords), novel words learned on the day of the fMRI scan (unconsolidated nonwords), and untrained nonwords (Davis et al., 2009). Presentation of untrained nonwords resulted in elevated hippocampal responses relative to unconsolidated nonwords, indicating that the hippocampus was involved in the formation of new phonological representations. In comparison, levels of cortical activity in the superior temporal gyrus were equivalent for unconsolidated and untrained nonwords, whereas the activation for consolidated nonwords was lower, closer to the level of activation for existing words. This suggests that novel phonological representations are integrated with pre-existing lexical knowledge in neocortical areas only after a period of sleep-associated offline consolidation.

One remaining question concerns the nature of representation in the two systems involved in memory and learning. Given the differences in the structure and function of the hippocampal and neocortical systems in the CLS framework it is feasible that these two systems are responsible for two different types of learning; learning of specifics and learning about generalities (O'Reilly & Norman, 2002). Whilst there is convincing evidence that the hippocampal system is important in episodic representation and memory, the qualitative nature of representation in the neocortical network is less clear since, as outlined in Chapter 1, there appears to be evidence that indexical, allophonic, and surface-form information are retained in long-term memory (Craik, 1991; Eisner & McQueen, 2006; Goldinger, 1996; Ju & Luce, 2006). Might consolidation processes strengthen and enhance memory for all information that is initially encoded and stored, or is consolidation a more selective process responsible for the generating of more abstract, context-free representations?

2.2.1 The hippocampal network

According to the CLS framework information is rapidly encoded into sparse representations in the hippocampal network, with a small number of highly selective units being used to represent each individual event. As such, each representation has minimal overlap with other representations, an essential requirement of episodic memory since it allows information about individual events to be kept separate from other similar events (O'Reilly, 2001).

The involvement of the hippocampal network in rapid encoding of novel lexical and semantic information has been demonstrated in a number of neuroimaging studies (Heckers, Weiss, Alpert, & Schacter, 2002; Mestres-Misse, Camara, Rodriguez-Fornells, Rotte, & Munte, 2008). For example, Breitenstein et al. (2005) showed that activation in the left hippocampus decreased as the number of form-meaning exposures increased during study, suggesting that the hippocampus was primarily responsible for rapid encoding of the items during initial exposures at the start of the learning phase. Interestingly, the magnitude of hippocampal activation observed during initial learning has been found to predict later memory for newly learned words (Breitenstein et al., 2005; Davis et al., 2009), suggesting that encoding of information within this system is vital in order for long-term memory representations to be later generated. In fact, when activity of the hippocampus is blocked (Riedel et al., 1999), or the hippocampus and related medial temporal lobe (MTL) structures are damaged (Alvarez, Zola-Morgan, & Squire, 1995; Penfield & Milner, 1958; Scoville & Milner, 1957; for a review see Winocur & Moscovitch, 2011) the ability to form new, long-lasting memories is disrupted. Nonetheless, there are a few intriguing studies demonstrating successful learning of new word-object associations following hippocampal damage in amnesic patients; Duff, Hengst, Tranel, and Cohen (2006) demonstrated that amnesic patients were able to learn self-generated labels for unknown Chinese tangrams, whilst Sharon, Moscovitch, and Gilboa (2011) showed rapid fast mapping of novel names and novel objects. Nevertheless, both studies failed to show learning when the task required learning of arbitrary mappings between names and objects using explicit encoding tasks. Thus, there is evidence that the hippocampus is involved in the acquisition of new memories under normal circumstances, although some types of learning may be achieved via alternative mechanisms when the hippocampal system is damaged. However, the extent to which these alternative mechanisms are involved in learning when the hippocampal system is intact remains unclear.

Evidence supporting the suggestion that hippocampal memories are likely to be sparsely represented comes from studies showing that the hippocampus has much sparser levels of firing than the neocortex (Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990; Boss, Turlejski, Stanfield, & Cowan, 1987) as well as an fMRI study by Bakker, Kirwan, Miller, and Stark (2008) demonstrating that when participants were exposed to pictures of the same object, but with small differences

in the image such as a change in object orientation, the CA3/dentate gyrus area of the MTL responded to the picture as though it was the first presentation of the object, indicating that even very small changes in the input can result in pattern separation within certain areas of the MTL. Moreover, the fact that a large number of cortical areas, as well as almost all association areas, have projections that converge in the hippocampus, makes this an ideal area to support the binding of information from different cortical areas to form highly-detailed episodic memory representations (Abutalebi et al., 2007; Mayes & Montaldi, 2001; Munoz & Insausti, 2005; Suzuki & Eichenbaum, 2000). In fact, patients with developmental amnesia, a syndrome in which there is selective damage to the hippocampus at birth or during early childhood, show severe deficits in episodic memory despite having intact semantic knowledge, and are typically unable to recall many details of their everyday life (Vargha-Khadem et al., 1997).

However, there is still debate as to how long the hippocampus remains involved in the processing of newly-encountered information. Some researchers claim that the hippocampus is only temporarily necessary for retrieval of newly acquired memories (*standard consolidation theories*; Squire & Alvarez, 1995; Squire & Zola, 1998; Teng & Squire, 1999; Zola-Morgan & Squire, 1990). One piece of evidence supporting this claim is that patients with hippocampal damage often show a temporally graded retrograde amnesia in addition to anterograde amnesia; that is to say that more recent memories appeared to be more severely disrupted than older memories (Nadel & Moscovitch, 1997). Nevertheless, findings of temporally graded retrograde amnesia spanning several years also suggest that the interaction between the hippocampus and neocortex is relatively prolonged, with a gradual reshaping of new memory representations that eventually results in memories being mediated by the neocortical system. This suggestion is consistent with work by Takashima et al. (2006) who examined the changes in neural correlates of memory for pictures of landscapes over 90 days. It was found that whilst the largest decrease in the amount of hippocampal activation occurred between Day 1 and Day 2, there were smaller, more progressive decreases in retrieval-related hippocampal activation across the whole 90 days, indicating that recall of novel information may continue to rely on hippocampal areas for an extended period of time after initial encoding. Thus, is seems reasonable to hypothesise that information retained within the highly-detailed, rapidly formed hippocampal representations may

continue to influence recognition of newly learned words for a considerable period of time following initial encoding.

Some researchers have even suggested that certain memories never become independent of the hippocampal system (Moscovitch, Nadel, Winocur, Gilboa, & Rosenbaum, 2006; Moscovitch et al., 2005; Nadel & Moscovitch, 1997; Winocur & Moscovitch, 2011). These researchers have proposed a *Multiple Trace Theory* (*MTT*), in which each individual experience is encoded as a unique memory trace consisting of a bound ensemble of hippocampal and neocortical neurons. Repeated exposure to the same item or event results in multiple memory traces that are distributed throughout the hippocampal formation. As a result, newly-encountered information is more susceptible to disruption as a result of hippocampal damage because traces associated with the novel information are not widely distributed. As such, MTT is able to account for the pattern of temporally graded retrograde amnesia that is typically assumed to support a standard consolidation account. Within MTT it is assumed that as long as episodic details about an item or event are available, the hippocampus will remain involved in memory (Rosenbaum, Winocur, & Moscovitch, 2001), as suggested by neuroimaging studies of autobiographical memory showing hippocampal activation regardless of the age of the memory (Addis, Moscovitch, Crawley, & McAndrews, 2004; Gilboa, Winocur, Grady, Hevenor, & Moscovitch, 2004).

Nonetheless, it is important to note that both standard consolidation theories and MTT assume that semantic memories (typically defined as general knowledge of the world) become gradually less reliant on the hippocampus and more reliant on the neocortex over time. The difference between these theories is their assumption about whether episodic memories (memory for autobiographical experiences and specific events) ever become independent of the hippocampus. Since vocabulary knowledge is typically considered as an aspect of semantic memory both theories would predict that memory for newly learned words should depend only temporarily on the hippocampal system. However this is not to say, particularly within the context of MMT, that specific details of previous encounters with the event must necessarily be lost as a result of the change in reliance on hippocampal and neocortical systems. Rather, it seems likely that access to and retrieval of highly-detailed hippocampal memories may depend on the task and on the salience of different aspects of the input during encoding.

2.2.2 The neocortical network

According to the CLS framework (McClelland et al., 1995) information is encoded slowly into the neocortical system allowing gradual integration of new and existing information as a result of the formation of overlapping, distributed representations that are sensitive to shared structure between different items. Establishment of neocortical representations requires a larger number of learning trials and/or periods of sleep-associated offline consolidation. Slow learning in this system is essential in order that new information can be integrated with existing information without existing knowledge being overwritten, a problem termed *catastrophic interference* within the connectionist modelling literature (French, 1999; McCloskey & Cohen, 1989).

Evidence of slow learning in the neocortex comes from neuroimaging studies indicating that neocortical contributions to memory increase over time while hippocampal contributions decrease (Bontempi, Laurent-Demir, Destrade, & Jaffard, 1999; Gais et al., 2007; Takashima et al., 2009; Takehara-Nishiuchi & McNaughton, 2008). For example, in the word learning study by Davis and colleagues (2009), described above, neocortical activation was observed only for novel words learned one day prior to the fMRI test session; in other words, evidence of neocortical representation emerged only after a period of sleep-associated offline consolidation. Takashima et al. (2009) have also demonstrated that connectivity between the hippocampus and neocortex decreases over time whilst connectivity within the neocortical network increases, consistent with the suggestion that neocortical representations become more dominant over time. Takashima and colleagues claim that "memories are gradually transferred to neocortical circuits with consolidation, where connections within this circuit grow stronger and reorganized so that redundant and/or contextual details may be lost" (2009, p.10087), a suggestion that, within the CLS framework at least, necessarily derives from the overlapping distributed nature of representations that are used within the neocortical system.

The assumption that abstraction occurs over time as memory becomes more reliant on neocortical regions is also consistent with MTT, which proposes a *transformation hypothesis* in which "the progression of memories from hippocampal to extra-hippocampal structures necessarily entails a loss of detailed, contextual features" (Winocur & Moscovitch, 2011, p.2). Note, MTT does not claim that the context-dependent memory is lost; rather it supposes that the generic memory

representation gradually becomes more dominant as time progresses, consistent with the assumptions of a CLS framework. One piece of evidence supporting a transformational hypothesis comes from a study by Winocur, Moscovitch, and Sekeres (2007) who examined context-dependent learning in rats. Shortly after study rats performed best in the same context as learning, but this same-context advantage disappeared when rats were tested after a longer delay suggesting that there is a transformation process involved in memory and learning that results in context-independent memory. In another study Winocur and Moscovitch examined memory for maze structure in rats with hippocampal and prefrontal lesions. Rats with hippocampal lesions showed disrupted memory for maze-specific information (*i.e.*, information about the maze in which they were trained prior to the lesion), but showed retention of maze-general search behaviour. In comparison, rats with lesions to prefrontal cortical regions showed deficits in maze-general, but not maze-specific behaviour. Together these findings support the suggestion that hippocampal and cortical regions are differentially involved in learning about specifics and learning about generalities (O'Reilly & Norman, 2002).

2.2.3 Complementary learning systems as a hybrid model of lexical representation

As outlined in Chapter 1 there are a number of lines of evidence pointing towards a hybrid model of lexical representation. A CLS account of memory and learning may offer one framework within which aspects of episodic and abstract representation could co-exist since hippocampal representations are likely to be highly-detailed in nature whilst neocortical representations may be more abstract. Interestingly Davis and Gaskell (2009) have suggested that activation of neocortical representations may occur more rapidly than activation of hippocampal representations in order that existing knowledge is activated prior to novel information, thus preventing catastrophic interference as new information is acquired. If this is the case then detailed indexical information retained within hippocampal representations will only be activated when sufficient processing time is allowed, thus accounting for the apparent slow integration of indexical information into the retrieved representations of existing word, consistent with Luce et al.'s (2003) time-course hypothesis.

Davis and Gaskell's suggestion gains support from a number of sources. Firstly, Feustel et al. (1983) found that identification of briefly presented words in a

repetition priming experiment was consistently faster for existing words compared to pseudowords, in line with the suggestion that access to episodic traces that are required for identification of newly-encountered pseudowords is slower than access to neocortical representations which presumably mediate identification of existing items. Secondly, in light of the apparent hemispheric differences in processing of abstract and episodic information, which are thought to occur preferentially in LH and RH respectively (Marsolek, 1999) Poeppel and colleagues suggest that the two hemispheres also differ in terms of their 'window of analysis'. The LH is thought to show greater involvement in early processing stages in the 20 to 50 ms window, and the RH showing greater involvement in later processing stages in the 150 to 300 ms window (Boemio, Fromm, Braun, & Poeppel, 2005; Poeppel, 2003), a suggestion that is consistent with Luce et al.'s (2003; McLennan & Luce, 2005) time-course hypothesis if we assume that the LH is preferentially involved in processing of abstract information and the RH in processing of episodic information. It is also interesting to note that Marsolek, Schacter and Nicholas (1996) propose that the two hemispheres are designed to be efficient at processing different types of information; LH may be more efficient at feature-based processing, which codes common features across variable inputs, whereas RH may be more efficient at whole-based processing that would be important for processing of highly-detailed episodic representations.

To summarise, the CLS framework offers one account of a hybrid model of lexical representation (Goldinger, 2007), with the hippocampal and neocortical networks accounting for two different types of learning; learning of specifics and learning about generalities respectively (O'Reilly, 2001; O'Reilly & Norman, 2002). These two types of learning require different neural architectures, with learning about specifics requiring a rapid learning rate and sparse representations, and learning about generalities requiring a slower learning rate to allow interleaving of old and new information into a system of overlapping, distributed representations (O'Reilly & Rudy, 2000). These two learning systems correspond well with Carey's (1978) notion that new lexical representation can be formed through fast mapping but that a more prolonged process of enhancement and stabilisation is required for a new word to be fully learned and integrated with existing lexical knowledge.

2.3 Aims and overview of thesis

The experiments reported in this thesis examine whether lexical representations are transformed qualitatively during the process of consolidation, as suggested by MTT, or whether extra-linguistic details are retained in long-term memory, as suggested by studies reporting surface-form specificity effects at delayed test points. Talker-specificity effects (TSEs) were examined over the course of a week in two tasks; recognition memory and tests of lexical competition. Assuming that recognition is primarily dependent on hippocampally mediated representations and that lexical competition effects are dependent on integration of novel representations with existing information in the neocortical network, it should be possible to use these tasks to determine the level of detail in representations within each system. Talker gender was selected as the indexical detail to be manipulated in these experiments since it is the most widely used variable in previous studies examining indexical-specificity effects for existing words. Previous studies have shown that changes in talker information between study and test result in slower and less accurate processing and recognition of recently studied existing words (Bradlow et al., 1999; Craik & Kirsner, 1974; Goh, 2005; Goldinger, 1996; Goldinger et al., 1999; McLennan & Luce, 2005; Schacter & Church, 1992; Sheffert, 1998).

Whilst there are a handful of experiments exploring the time-course of surface-form specificity effects for existing word (Craik, 1991; Eisner & McQueen, 2006; Goldinger, 1996; Ju & Luce, 2006), the time-course of these effects has not yet been examined for newly-learned words. Experiment 1 aimed to replicate the finding that TSEs are present immediately after novel words have been studied, consistent with previous research (Creel et al., 2008; Creel & Tumlin, 2009, 2011). Moreover, we attempted to demonstrate the presence of TSEs in three different tasks in order to ensure that these effects were not task-specific. Experiments 2 and 3 explored changes in TSEs in recognition memory and tests of lexical competition for novel words over the course of a week. If consolidation processes strengthen memory for all types of form-based information, as has already been demonstrated for phonological information (*e.g.*, Dumay & Gaskell, 2007), then larger TSEs may be expected after a period of sleep-associated offline consolidation, alongside significant talker-specific lexical competition effects. Together this would provide support for exemplar theories of lexical representation. Alternatively, if consolidation is a more selective process, as suggested by the transformation

hypothesis, then TSEs should decrease over time as talker-general lexical competition effects strengthen. Alternatively, if episodic representations are maintained even after abstract representations are generated and stabilised then the time-course of TSEs and lexical competition effects may be independent. Either of these latter two findings would provide support for a hybrid model of lexical representation.

Chapter 4 describes two experiments directly comparing the time-course of TSEs in recognition memory for existing and novel words, controlling the number of talkers, number of each type of item, number of exposures to each item, and number of test-points completed by each participant. These experiments were included to ensure that the contrasting time-courses of TSEs observed for existing words in Goldinger's (1996) study, where TSEs decreased over a week in an old/new recognition task, and those for novel words in Experiments 1 to 3 where TSEs were significant at all time points, were not simply due to methodological differences.

Chapter 5 offers a systematic review of the effects of within- and between-talker variability on the time-course of TSEs for novel words. The purpose of this manipulation was to try to mimic, in a systematic manner, the type of exposure that people have to existing words where items are likely to be encountered in multiple voices, at multiple speech rates, and spoken with different stress and intonation patterns. Chapter 5 also contains cross-experiment analysis examining the differences between TSEs for novel words trained with no variability, within-talker variability, and between-talker variability in both recognition memory and in terms of lexical competition effects. Finally, the general discussion attempts to draw together the complex set of results, and evaluates the degree to which episodic and abstract representations may be involved in spoken word recognition at different time points during word learning.

CHAPTER 3: TALKER-SPECIFICITY AND LEXICAL COMPETITION EFFECTS DURING WORD LEARNING

3.1 Introduction

The three experiments reported in this chapter examine whether information about the study talker is encoded and stored when a new word is initially encountered, and whether this information affects later recognition of the new word. Experiments 2 and 3 also explore the time-course with which the same novel words begin to engage in lexical competition with phonologically similar existing words, as well as the possibility that talker-specific information may influence these lexical competition measures. If we assume (within a CLS framework, McClelland et al., 1995) that recognition of newly-learned items depends primarily on hippocampally-mediated representations, but that lexical competition effects are more dependent on integration of novel and existing information in the neocortical system, then it remains possible that these two effects rely on different types of representations[2]; presumably TSEs are dependent on the availability of highly-detailed episodic representations, whilst lexical competition is driven by more abstract phonological representations. If this is indeed the case, then the question remains as to whether the time-course with which the two types of representation are established is the same or different, and whether the two types of representation can co-exist in memory or whether the emergence of more abstract representations is a driving factor in the decay or loss of episodic information from memory.

To date the time-course of TSEs and lexical competition effects have not been examined using the same stimuli. Studies examining the emergence of lexical competition effects for newly-learned words (outlined in more detail in Chapter 2) suggest that novel and existing items are integrated into the existing lexicon only after a period of sleep-associated offline consolidation (*e.g.,* Dumay & Gaskell, 2007). This finding indicates that the representations that are necessary for lexical competition effects are not sufficiently robust to support these effects immediately after study, but are strengthened over time. Contrary to the pattern of lexical

[2] Note that the experiments reported in this thesis are all behavioural. The CLS framework (McClelland et al., 1995), in which hippocampal and neocortical networks may underlie systems of epidosic and abstract representation (respectively) is used only as an analogy to the possible neural mechanisms underlying the behavioural effects observed. Further experiments involving neuroimaging techniques are required in order to verify the claims made about different types of lexical representation in these two neural systems (see section 6.3.3).

competition data, a number of studies examining TSEs for existing words have indicated that talker information affects recognition of existing words immediately after they have been studied (see Chapter 1 for a review). Together these two sets of data provide preliminary evidence that the type of representations used to support lexical competition effects and TSEs may differ.

Nonetheless, despite the relatively large number of studies examining TSEs for existing words, few have examined TSEs during the recognition of newly-learned words. In one study examining TSEs for novel words Creel et al. (2008) taught participants to identify novel nonwords associated with novel objects. During study each novel word was heard in only one voice. Critically, the target and competitor items (either a novel cohort or rhyme competitor) were either spoken consistently by the same talker, or consistently by different talkers. During the test trials more fixations to the target item and fewer fixations to the competitor item were observed when the target and competitor had been spoken by different talkers during study. These findings demonstrate that talker-specific information is encoded and stored in memory for novel lexical items, and that this information affects the degree to which two phonologically-similar novel words engage in lexical competition (see Creel & Tumlin, 2009; 2011, for similar findings). However, in Creel et al.'s experiment participants were required to learn pairs of phonologically-similar novel nonwords (*e.g.*, *aruju-aruja*), and to associate each novel item with a complex abstract shape. As such, talker information may have been deliberately encoded as part of a conscious strategy to aid learning, particularly in the case where the two phonologically-similar nonwords were spoken by different talkers. Moreover, Creel et al.'s study demonstrates only that talker information affects lexical competition within a small set of novel words; they did not investigate whether talker information can affect the amount of lexical competition that is observed between existing and novel words. Nevertheless, their findings indicate that extra-linguistic details are encoded and stored in memory when novel words are initially encountered, just as they are for existing words. Further evidence supporting this suggestion comes from an experiment by Brown and Carr (1993) demonstrating faster naming and lexical decision responses to pseudowords when presented in the same visual form (*e.g.*, handwritten - handwritten) at both study and test compared to when surface-form details changed between study and test (*e.g.*, handwritten - typeface). Both of these findings are consistent with the idea that representations of new words are episodic

in nature and that these episodic representations are established rapidly when a novel item is encountered, consistent with predictions of a CLS framework (McClelland et al., 1995).

However, few studies have examined the time-course of surface-form specificity effects. Goldinger (1996) examined the time-course of TSEs for existing words, but found different patterns of data in two different tasks. Participants were exposed to 150 monosyllabic existing words spoken by two, six, or ten talkers, and were required to type the word that they had heard during each study trial of the experiment. At test participants completed an old/new categorisation task, in which the 150 studied items were heard alongside 150 filler words, with half of the studied items spoken by the same talker as study, and half spoken by a different talker. Participants were also required to identify the words when they were heard in a background of white noise. These two test tasks were completed either five minutes after the study phase, one day later, or one week later. In the old/new categorisation task TSEs were observed only after delays of five minutes and one day, not one week later. By comparison the identification-in-noise task revealed significant TSEs at all delays, although even in this task the size of the same-talker advantage decreased significantly over time despite the fact that TSEs remained significant on Day 8. Thus, whilst talker-specific information can affect processing of lexical items one week after they are initially encountered in some tasks, it also appears that this information becomes gradually less useful in others as time progresses. Nonetheless, the presence of significant TSEs at all time-points in the identification-in-noise task (despite the significant decrease in the size of the TSEs over the same time period) suggests that representations containing detailed talker information must be maintained for at least a week after studying an existing word.

Goldinger's (1996) observation that TSEs decreased over time for existing words in his old/new categorisation task stands in contrast to the observation that lexical competition effects between existing and novel words are typically absent immediately after a novel nonword is studied, but emerge one day later (*e.g.*, Gaskell & Dumay, 2003). The different time-courses of lexical competition effects and TSEs in recognition memory observed in previous studies suggest that as the representations in the neocortical system that are assumed to be important in allowing competition to occur between similar-sounding items become more dominant during the processing and recognition of spoken words, episodic details

may be lost from memory or at least may become less influential. However, as noted in Chapter 1, using existing words to examine the time-course of TSEs is problematic since participants will have heard the existing words spoken many times prior to an experiment, in many different voices, and in varying speech rates, amplitudes, and intonations. Thus, the decline in TSEs in Goldinger's old/new categorisation task could be explained by both hybrid and episodic models of lexical representation. According to a hybrid model, the neocortical system would contain only canonical phonological information and thus talker information should not affect recognition of items once the neocortical system becomes dominant. Alternatively, episodic models may assume that episodic traces associated with each encounter with an existing word are stored in memory and that multiple traces are partially activated when that existing word is heard. As a result, the effects of talker information on processing of existing words are minimised once multiple traces are activated at retrieval. Thus, it remains possible that representations in the neocortical system may be either abstract or episodic in nature.

The three experiments reported in this chapter examine the time-course of TSEs and lexical competition effects concurrently for the same set of novel nonwords. The aim of these experiments was to provide clearer information about the nature of representations within the neocortical network. If representations in the neocortical system are episodic in nature then lexical competition effects should be talker-specific, and recognition of novel words should remain highly talker-specific at all time points. On the other hand, if representations in the neocortical network are abstract in nature, consistent with McClelland and colleagues' (1995) suggestion that the representations within this system are overlapping and distributed in nature, then lexical competition effects should not be affected by information about the study talker of a novel word, and there may also be a decrease in TSEs for novel words as the neocortical system becomes more dominant over time. However, if highly-detailed hippocampal representations and abstract neocortical representations are able to co-exist, and the establishment of abstract representations is not responsible for driving the decay or loss of episodic details, then the time-course of lexical competition and TSEs may differ. More specifically, given evidence from patients with temporally-graded retrograde amnesia (Nadel & Moscovitch, 1997), it seems plausible that episodic representations maintained by the hippocampal system may be long-lasting, and may continue to contribute to recognition of newly-learned

words under certain conditions even after abstract representations in the neocortical system have been firmly established.

In all three experiments participants were exposed to 24 novel nonwords in a study task, with half of the items consistently spoken by a male talker and half consistently spoken by a female. During all test tasks half of the studied items changed talker and half remained in the same voice. Experiment 1 examined TSEs immediately after study in three different tasks in order to select a suitably robust task for use in later experiments examining the time-course of TSEs. Experiments 2 and 3 explored the time-course of both TSEs and lexical competition effects over the course of a week for the same set of 24 novel words. In order to investigate the emergence of lexical competition in Experiments 2 and 3 it was necessary to compare response times (RTs) from a list of basewords with novel nonword competitors (test items) and RTs to basewords without novel competitors (control items). To ensure that TSEs were observed immediately after study for all stimuli that were used in Experiments 2 and 3, both stimulus lists were included in Experiment 1, counterbalanced across participants as a between-participants variable.

3.2 Experiment 1

Experiment 1 aimed to investigate whether talker-specific information was retained in memory for novel words immediately after study when only the phonological form of the novel word was provided and all novel words were phonologically dissimilar. A second goal of Experiment 1 was to select a suitably robust task that could be used in later experiments to examine TSEs in memory for novel words at multiple time points. Although acquisition of phonological information is only one of many facets of lexical acquisition it is a fundamental part of word learning; for a spoken word to be recognised it must find a matching representation in the lexicon irrespective of semantic and orthographic information.

Participants were exposed to 24 novel nonwords (*e.g. biscal*) in a phoneme monitoring study task. Following a short maths-based distracter task, included in order to minimise short-term recency effects in later recognition tasks (Arnon & Ramscar, 2009; Goh, 2005), there were two experimental tasks designed to tap memory for talker-specific information. First was a stem-completion task, a task that has previously been used to demonstrate TSEs in memory for existing words

(Church & Schacter, 1994; Schacter & Church, 1992). The second task combined two previously used tests of TSEs. Participants heard a list of nonwords, half of which had been studied during phoneme monitoring (*e.g., biscal*), the other half being phonologically-similar foil words (*e.g., biscan*), and were required to make a decision as to whether the spoken nonword had been heard previously during the study phase of the experiment (*old*) or had never been heard before (*new*). Following a button-press response in answer to this question participants were cued to repeat back the nonword that they had heard at the start of the trial, thus incorporating a delayed shadowing task with an old/new categorisation task.

One criticism of old/new categorisation tasks that use existing words is that even when the listener must respond '*new*', indicating that an item has not previously been encountered within the experimental session, participants will have encountered those '*new'* words during everyday life. Thus, the retrieval and use of talker-specific information may be part of a deliberate or intentional strategy used to help participants to remember which existing words have been previously encountered within a specific experimental list or session. The use of novel nonwords in our old/new categorisation task avoids this problem; participants had never before heard the '*new*' foil nonwords.

During the study phase of the experiment half of the novel nonwords were spoken consistently by a male talker, and the other half consistently by a female talker. At test half of the items spoken by each talker changed to the opposite talker whilst the other half remained in the same voice. The voices remained constant across the testing phase such that an item heard in the male voice in the stem-completion task was also heard in the male voice in the recognition with delayed shadowing task, and the same for items heard in the female voice. It was predicted that if detailed talker-specific information was encoded and stored when novel lexical items were encountered then changes in talker between study and test would result in poorer performance in all of the tests of TSEs. Such a finding would support data from studies by Creel and colleagues (Creel et al., 2008; Creel & Tumlin, 2009, 2011), as well as the suggestion that hippocampally mediated representations that are formed immediately upon encountering novel lexical items are indeed highly detailed, containing information beyond the basic canonical phonemic representation of the novel items.

3.2.1 Method

Participants

Thirty-two undergraduate students (*age range* = 18–26 years, 6 male) from the University of York participated in the experiment and were rewarded with either payment or partial course-credit. Participants in this, and all subsequent experiments, were native speakers of British English and reported no known hearing, speech or language impairments at the time of testing. Informed consent was obtained from all participants prior to the first session.

Stimuli

Forty-eight stimulus triplets, each containing one existing baseword and two novel words, were selected from stimuli used by Tamminen and Gaskell (2008) in a longitudinal study of word learning in adults. All basewords were monomorphemic and had uniqueness points located at or before the final vowel. The novel words differed from their baseword at the final vowel (*e.g. biscuit /*biski***t**/ and *biscal /*biskə***l**/), and from each other at the final consonant or consonant cluster (*e.g. biscal /*biskə***l**/ and *biscan /*biskə***n**/). All three words were produced using the same stress pattern. Throughout this thesis the novel items encountered during the study phase of the experiments will be referred to as *novel nonwords* where as the untrained novel items that are used as distracters in the old/new categorisation task will be referred to as *foil nonwords*.

Stimulus triplets were selected such that two lists of 24 basewords, matched on initial phoneme and number of syllables (12 bisyllabic and 12 trisyllabic per list), could be created. The two lists were matched as closely as possible in the number of phonemes (M = 7.96, *Range* = 6 - 11) and frequency (M = 3.63, *Range* = 2 - 14) according to the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993). Independent samples t-tests showed that there were no significant differences between the two lists in either number of phonemes, t(46) = .211, *ns*, or frequency, t(46) = .116, *ns*. All basewords, novel nonwords, foil nonwords and their corresponding properties are listed in Appendix A.

The stimuli were recorded in a sound attenuated booth by one male and one female talker, both native British English speakers, using a Marantz CD recorder and Sennheiser ME40 microphone. The stimuli were digitized at a 44.1Hz sampling rate with 16-bit analogue-to-digital conversion. Adobe Audition was used to normalize

the peak amplitude of all stimuli and to generate stem-cues incorporating the first syllable (CVC) of each novel word for the stem-completion task.

On average, stimuli spoken by the male talker were slightly shorter than those spoken by the female talker. Paired samples t-tests indicated that this difference was significant for all sets of stimuli (basewords – $t(49) = 14.980$, $p < .001$; novel nonwords – $t(49) = 11.455$, $p < .001$; foil nonwords – $t(49) = 8.981$, $p < .001$; stem-cues – $t(49) = 5.478$, $p < .001$). Although this difference in articulation rate between talkers was unplanned, and will have added to the indexical differences between talkers, this is not greatly important since Experiments 1 to 3 were primarily interested in the time-course of TSEs for novel words, rather than in the specific variables driving the TSEs themselves.

Design

Participants were tested individually in a sound-attenuated room. Tasks were run on a Carrera SSC computer using DMDX experimental software (Forster & Forster, 2003). Stimuli were presented binaurally over Beyerdynamic DT 294 headphones at a comfortable listening level. Button-press responses were made using an 850F Vibraforce Feedback Sightfighter game-pad and were recorded by DMDX, with RTs measured from stimulus-onset. Verbal responses were recorded using a head-mounted microphone. RTs for verbal responses were measured up to the onset of the voice key trigger. The same equipment was used to run Experiments 2, 3, 4, and 5.

All tasks in Experiment 1 were completed during a single session lasting 45-50 min. Participants were allowed to take breaks between tasks and between blocks in the phoneme monitoring task in order to maintain their concentration and attention throughout the session.

Procedure

During the *study phase* of the experiment each participant was exposed to one list of 24 novel nonwords, counterbalanced across participants, in a phoneme monitoring task. Within this list 12 items were spoken consistently by the male talker and 12 consistently by the female talker.

Participants listened for specified phonemes in the novel nonwords, indicating the presence or absence of the target phoneme through a button-press response. The

task began with five existing-word practice trials, followed by six experimental blocks, each specifying a different target phoneme. For all participants these blocks occurred in a fixed order (/p/, /t/, /b/, /m/, /s/, /d/) with the novel nonwords occurring three times in each block. The order of the novel items was randomised in groups of 24 (*i.e.*, one full repetition of the novel nonword list) in order to reduce the chance of the same item occurring twice in a row or in close proximity within blocks as this may have drawn attention to the fact that the same items were always spoken by the same talker during study. Target phonemes occurred at all positions across the novel nonwords, with the number of target present trials varying both between lists and between blocks.

Throughout each block of phoneme monitoring the target phoneme was displayed centrally on the computer monitor and a tick and cross were displayed in the bottom left and right corners of the screen respectively, above the appropriate response keys. Instructions emphasised that responses should be made both quickly and accurately. At the end of each block of phoneme monitoring participants were provided with feedback about their average RT for that block and the number of errors made. This feedback was included in order to encourage participants to continue responding as quickly and accurately as possible throughout the task. All RTs were measured from word onset, with a maximum RT of 5s, after which the program automatically moved on to the next item with an inter-trial interval of 500ms.

Following the phoneme monitoring task participants completed a short *distracter task*, a pen-and-paper maths verification task in which they had to indicate whether 24 simple sums (*e.g.* "(8 x 2) + 3 = 20") were correct or not by circling the word "correct" or "incorrect" on a response sheet provided. On average participants took 205s (*SD* = 65s) to complete the maths task. Data from this task are not reported.

During the *test-phase* participants completed two tasks; stem-completion and an old/new categorisation task with delayed shadowing. For these tasks the lists of items heard during the study phase were subdivided once more so that half of the nonwords stayed in the same voice as study, and half changed. In other words, six of the 12 nonwords heard in the male voice during study were heard in the female voice at test, whilst the other six remained in the male voice, and likewise for words heard in the female voice during study. Overall, at test, 12 nonwords were heard in the

same voice as at study, and 12 in a different voice. The counterbalancing of talker remained constant across both of the test tasks. In order to account for the fact that one exposure to the novel nonwords or first syllable of the novel nonwords in a different voice to exposure during the first test task may have impacted upon performance in the second task, the order of the two tasks was counterbalanced. This manipulation allowed us to determine the robustness of TSEs in memory for novel nonwords immediately after study. Instructions emphasised speed and accuracy in all test tasks. In order to avoid drawing attention to our same-talker/different-talker manipulation participants were not informed that half of the items would change talker between study and test in this experiment.

During the *stem-completion* task participants heard the first syllable (CVC) of the novel nonwords to which they had been exposed during study and were required to complete these word-stems. Each trial began with a central fixation cross (+) displayed on screen for 500ms, followed by a delay of 500ms before the word-stem cue was played. After hearing the word-stem participants were required to say the novel nonword that completed this stem. Instructions emphasised that responses should only include words heard in the phoneme monitoring task. The maximum RT, measured from the onset of the stem-cue, was 5s, after which the program automatically moved on to the next item. However, once a response was made (as determined by DigitalVOX calibration) 1s of audio response was recorded before moving to the next item.

The second test task, *old/new categorisation with delayed shadowing*, incorporated two types of test that have previously been used to look at episodic effects with existing words. These two tasks were combined in order to reduce the number of exposures to each novel nonword during test, and thus minimize carryover effects from hearing the voice change between study and test since TSEs may be smaller in the second experimental test task as a function of participants having heard half of the words change voice in the first test task.

As in the stem-completion task each trial began with a central fixation cross (+) displayed on screen for 500ms, followed by the words '*old*' and '*new*' displayed on the left and right sides of the screen respectively. After 500ms either a novel nonword (*e.g., biscal*) or a foil word (*e.g., biscan*) was heard and participants were required to decide whether the item was old (heard during the phoneme monitoring task) or new (had never been heard before). RTs were recorded from word onset

until a button-press response was made. The words '*old*' and '*new*' remained on screen until a response had been made, after which they disappeared and were replaced with a blank screen for 1s, followed by a star (*) presented centrally on screen cueing participants to repeat the nonword that they heard at the beginning of the trial. Participants had up to 2.5s to respond before the program moved on to the next trial. Spoken responses were recorded by DMDX as in the stem-completion task. Trials were separated by 1s.

3.2.2 Results

For all analyses in this and all subsequent experiments word list (1 *vs.* 2) was included as a dummy variable in order to reduce the estimate of random variation (Pollatsek & Well, 1995). Significant main effects and interactions involving this variable are reported only for the study task. In addition, all percentage correct scores were subject to an arcsin(√) transform in order to better meet the assumptions of normality.

Study phase

Sixteen participants were exposed to List 1 (2 male) and 16 to List 2 (5 male). Within each list, 8 participants completed the stem-completion task first (List 1 = 1 male; List 2 = 2 male) and 8 completed the recognition with delayed shadowing task first (List 1 = 1 male; List 2 = 3 male). The mean error rate in the phoneme monitoring task was 5.89% (*SD* = 2.83%), indicating that participants paid close attention to the phonological form of the novel nonwords during study. A repeated-measures ANOVA, with factors study talker (male *vs.* female), task-order (stem-completion first *vs.* categorisation with delayed shadowing first), and list (1 *vs.* 2), showed that there were no significant main effects of these variables on error rate (talker – $F_1(1,28) = 2.51$, *ns*, $F_2 < 1$; task order – $F < 1$; list – $F < 1$, nor were there any significant interactions between these variables. Therefore, any differences in memory for the novel nonwords in the test tasks are unlikely to be due to differences in performance during study for different lists and/or talkers.

It is important to note that RTs in the phoneme monitoring task differed significantly as a function of study talker (male: *M* = 1090ms, *SD* = 230ms; female: *M* = 1179ms, *SD* = 266ms), $F_1(1,28) = 37.88$, $p < .001$, $\eta_p^2 = .58$, $F_2(1,46) = 110.96$,

$p < .001$, $\eta_p^2 = .71$. The most likely explanation for the main effect of study talker is that the female tokens of the novel words were on average longer than the male tokens. RTs in the phoneme monitoring task are likely to reflect this difference since phonological information would unfold more slowly in the female tokens of the novel words than in the male tokens, and thus the relevant information that would allow participants to respond to the item would have occurred later in the female tokens. RTs did not differ between list, $F_1(1,28) = 1.75$, *ns*, or different task orders, $F_1 < 1$, in the by-participants analysis, although both of these variables showed significant main effects in the by-items analysis (list – $F_2(1,46) = 39.83$, $p < .001$, $\eta_p^2 = .46$; task order – $F_2(1,46) = 8.28$, $p < .01$, $\eta_p^2 = .15$), suggesting participants exposed to List 1, and participants who were later given the stem completion task first responder faster in the phoneme monitoring task than participants exposed to List 2, or those given the old/new categorisation with delayed shadowing task first. However, given the differing results between by-participants and by-items analyses for these two variables it would be unwise to place too much emphasis on these data.

Talker-specificity effects

In the *stem-completion task* verbal responses were scored according to the accuracy of the final syllable of the novel nonwords since the first syllable was always provided as the cue, and in all cases the second syllable of trisyllabic nonwords was identical to that of the existing baseword. No points were awarded if the participant did not respond (25.91% of trials), if the first one or two syllables were incorrect, or if the final syllable was completely incorrect. One point was awarded if either the final consonant or final vowel was correct, and two points were awarded where both the final vowel and final consonant were correct. For each participant the total score was converted to a percentage by dividing by 48 (2 points x 24 items) and multiplying by 100. Mean accuracy in stem-completion was 39.26% (*SD* = 19.43%).

RTs were measured from the onset of the cue, and were determined for each item using CheckVocal software (Protopapas, 2007). Prior to analysis all RTs corresponding to incorrect responses and RTs under 300ms were removed from the data set. No upper cut-off point was used for RT data since the accuracy scores indicated that the stem-completion task was difficult, resulting in lots of missing data

cells from incorrect responses alone. Mean RT was 1383ms (*SD* = 474ms). Table 3.1 shows accuracy (%) and RT (ms) in the stem-completion task as a function of whether the word stem was heard in the same or a different voice as compared to the exposure phase.

Repeated-measures ANOVAs were conducted separately for accuracy and RT data, with variables test-phase talker (same *vs.* different to study), and task order (stem-completion first *vs.* categorisation with delayed shadowing first). Analysis of the accuracy data revealed a significant main effect of test-phase talker, $F_1(1,28) = 4.97$, $p < .05$, $\eta_p^2 = .15$, $F_2(1,46) = 3.07$, $p = .09$, $\eta_p^2 = .06$, with participants showing greater accuracy in recall of items spoken in the same voice at study and test compared to items spoken in a different voice. There was no main effect of task order, $F_1(1,28) = 1.61$, *ns*, $F_2(1,46) = 9.07$, $p < .01$, $\eta_p^2 = .17$, nor were there any significant interactions. For the RT data there were no significant main effects of test-phase talker, $F_1(1,28) = 1.57$, *ns*, $F_2 < 1$, or task-order, $F_1 < 1$, $F_2(1,32) = 1.47$, *ns*, nor were there any significant interactions.

Additional post-hoc analyses were carried out in order to determine whether the study voice (male *vs.* female) and/or the test voice (male *vs.* female) affected the data. These analyses were conducted for all of the tests of TSEs in this and all subsequent experiments, and are reported in Appendix B.

Table 3.1. Accuracy and RT measures in the stem-completion, old/new categorisation, and shadowing tasks reported as a function of whether the novel word was heard in the same voice or in a different voice to the study phase of the experiment. *NW* – novel nonword, *FO* – foil nonword**.**

Task	Measure	Same talker	Different talker
Stem-completion	NW accuracy (%)	42.2	36.3
	NW RT (ms)	1326	1398
Old/new categorisation	d′ (SDT)	2.31	1.57
	β (SDT)	0.74	2.15
Shadowing	NW accuracy (%)	82.9	86.8
	FO accuracy (%)	86.4	87.4

In the *old/new categorisation task* all data points corresponding to incorrect responses were removed prior to analysis. In addition, RTs more than 2.5 *SD* above or below the mean RT for each individual participant were removed. One participant and two items had error scores more than 2.5 *SD* above the grand mean in the by-

participants and by-items analyses respectively and were removed from the data set. Mean accuracy was 77.5% (*SD* = 7.4%). RTs were measured from word onset until participants made an old/new categorisation button-press response. The mean RT was 1441ms (*SD* = 228ms).

Data from the old/new categorisation task were analysed using signal detection theory (*SDT*; Green & Swets, 1966). SDT is used to get an estimate of sensitivity (the ability to distinguish signal from noise) that is unaffected by individual differences in response bias. Two measures were calculated; d-prime (d′) and beta (β). D′ provides a measure of sensitivity; this is a measure of how well participants were able to discriminate between old/studied and new/unstudied items. Low d′ scores indicate that participants were performing close to chance, and were unable to differentiate between studied and unstudied items, whereas higher d′ scores indicate that participants were more successful at discriminating studied from unstudied items. Although d′ scores are the most important SDT measure in the current studies, β scores are also reported to provide an indication of the extent to which participants altered their response criterion depending on whether the item was heard in the same or a different voice to study. Values of one indicate that participants were not biased towards either '*yes/old*' or '*no/new*' responses, values less than one indicate a bias towards '*yes/old*' responses, and values above one indicate a bias towards making '*no/new*' responses.

To calculate d' and β the number of hits, misses, false alarms, and correct rejections (Table 3.2) were calculated for same-talker and different-talker items separately. Hit rates (*H*; proportion of 'studied' trials on which participants responded 'old') and false alarm rates (*F*; proportion of 'unstudied' trials on which participants responded 'old') were calculated. These scores were then used to calculate d′ (Formula 3.1) and β (Formula 3.2). Repeated-measures ANOVAs were conducted separately for d′ and β data, with the same factors included in the analysis as above.

Table 3.2. The four possible types of response in signal detection theory

	Response: OLD	**Response: NEW**
Stimulus: STUDIED	Hit	Miss
Stimulus: UNSTUDIED	False alarm	Correct rejection

Formula 3.1. $d' = z(H) - z(F)$

Formula 3.2. $\beta = EXP((z(F)^2 - z(H)^2)/2)$

For the d′ data there was a significant main effect of test-phase talker, $F(1,27) = 21.54$, $p < .001$, $\eta_p^2 = .44$, indicating that items were categorised more accurately when heard in the same voice as study (see Table 3.1). There was also a significant main effect of test-phase talker for the β values, $F(1,27) = 45.86$, $p < .001$, $\eta_p^2 = .63$, suggesting that for same-talker items participants were biased towards responding '*old*', whilst for different-talker items participants were biased towards a '*new* response. The main effect of task-order was not significant in either analysis. Analysis of RT data for this and all subsequent old/new categorisation tasks is reported in Appendix C.

In the *delayed shadowing task* only accuracy data were analysed. This was due to a technical error that resulted in inaccurate recording of the speech onsets. As a result RT data could not be included in the analysis. Shadowing responses were scored for accuracy, receiving two points for completely correct words, one point for words containing only one phonemic error, and zero points for words containing two or more phonemic errors. Overall participants correctly shadowed the novel and foil words 85.25% of the time (*SD* = 7.39%). Data from one participant was excluded prior to analysis due to having an accuracy score more than 2.5 standard deviations below the mean. Table 3.1 shows accuracy scores (%) for both same-talker and different-talker items.

A repeated-measures ANOVA for the accuracy data, with factors as above, revealed that there was a main effects of test-phase talker, $F_1(1,27) = 4.52$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,92) = 4.39$, $p < .05$, $\eta_p^2 = .05$, with participants showing poorer shadowing responses when items were heard in the same voice as study, contrary to predictions. Further analysis revealed that this main effect of test-phase talker was marginally significant for the novel nonwords, $F_1(1,27) = 3.40$, $p = .076$, $\eta_p^2 = .11$, $F_2(1,46) = 5.44$, $p < .05$, $\eta_p^2 = .11$, but was non-significant for the foil nonwords, $F < 1$, indicating that talker information influenced repetition of studied items only. No other main effects or interactions approached significance.

3.2.3 Discussion

Data from Experiment 1 show that accurate phonological representations are formed immediately upon encountering novel nonwords, consistent with previous studies examining word learning in adults (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008; Tamminen et al., 2010). In addition, Experiment 1 demonstrates that detailed talker-specific information is retained in memory when novel words are initially encountered, and that this information can affect recognition several minutes later, consistent with eye-tracking studies by Creel and colleagues (Creel et al., 2008; Creel & Tumlin, 2009, 2011). Just like existing words (Church & Schacter, 1994; Schacter & Church, 1992), novel nonwords were recalled more accurately when the stem-cue was heard in the same voice as study. Likewise, novel nonwords were categorised as '*old*', previously-encountered nonwords more accurately when the novel item was heard in the same voice that it had originally been encountered in, again, consistent with studies using existing words (Bradlow et al., 1999; Goh, 2005; Goldinger, 1998; Palmeri et al., 1993; Pilotti et al., 2000; Sheffert, 1998). Although it could be argued that the same-talker advantage in old/new categorisation may have arisen due to participants responding '*old*' only in cases where they thought the item was both a studied item *and* heard in the same voice as study (since participants were not explicitly told to ignore talker information and focus only on recognition of the phonological form of novel words during this task), data presented later in Experiment 3 argue against this possibility since participants in this experiment were explicitly told to ignore information about the test-talker used in the old/new categorisation task.

It is interesting that a same-talker advantage was not observed in the delayed shadowing task. Previous research using existing words has demonstrated that TSEs are found in delayed shadowing, but not in immediate shadowing (McLennan & Luce, 2005). McLennan and Luce suggest that the delay between hearing an item and shadowing it provides additional time to integrate additional extra-linguistic details into the retrieved representation. One possible explanation for the lack of a same-talker advantage in the current experiment may be that requiring participants to make an additional old/new decision during the delay between hearing a word and shadowing it makes it more difficult to integrate indexical and phonological information in a way that will aid production of that item, although the presence of

TSEs in the old/new categorisation task, which participants completed prior to shadowing the words, suggests that phonological and indexical information are likely to have already been integrated.

Despite the null effects in our delayed shadowing task Experiment 1 was able to address whether TSEs are robust for novel words immediately after study. Since no main effects of task-order were observed in any of the analyses it appears that hearing a word in two voices, one during study, and one during the first test task, did not eliminate TSEs in a second test task. One reason for the lack of interference caused by hearing a novel word in a second voice during test may be that sparsely coded hippocampally-mediated representations were formed for the novel word when heard in each voice. The non-overlapping nature of representations in the hippocampus would then minimise interference between these stimuli.

Given that Experiment 1 provides evidence that talker-specific information is encoded and stored when novel words are initially encountered, Experiments 2 and 3 examine the time-course of these specificity effects in recognition of novel nonwords over the course of a week, as well as exploring whether talker information affects lexical competition between the novel nonwords and similar-sounding existing word.

3.3 Experiment 2

The old/new categorisation task was selected from Experiment 1 as the test of TSEs for Experiments 2 and 3 since it showed a robust same-talker advantage in both accuracy and RT analyses. An additional advantage of the old/new categorisation task is that, unlike the stem-completion and shadowing tasks, accuracy scores did not appear to be influenced by which voice an item was heard in during either study or test (see Appendix C). Moreover, old/new categorisation does not require participants to articulate the novel items. One concern in Experiments 2 and 3, where participants were tested at multiple time-points, was that requiring participants to articulate the novel nonwords might result in an additional representation of each of the novel nonwords in the participants' own voice, which may affect measures of TSEs at later time points.

In Experiment 2, participants were again exposed to one list of 24 novel nonwords in a phoneme monitoring task, during which half of the novel nonwords were heard consistently in a male voice and the other half consistently in a female voice. At test participants completed one test of TSEs (old/new categorisation) and

one test of lexical competition (lexical decision). In the test of lexical competition participants were required to make speeded lexical decisions (*word*/*nonword*) to auditory stimuli. According to the cohort model of spoken word recognition (Marslen-Wilson & Zwitserlood, 1989) lexical competition between phonologically-similar words occurs up to the point at which only one word in the lexicon matches the speech input (the uniqueness point of the word). By teaching participants novel nonwords (*e.g., biscal*) that differ from their basewords (*e.g., biscuit*) only after the theoretical uniqueness point of that word, it is possible to artificially shift the uniqueness point towards the offset of the basewords. As a result, slowed processing of the basewords in a lexical decision task would indicate that the baseword and novel nonword were engaging in lexical competition.

In the lexical decision task participants heard the 24 basewords from which the 24 studied novel nonword had been derived (*test basewords*), as well as 24 basewords from which the unstudied list of 24 novel nonwords had been derived (*control basewords*). Test basewords were spoken by the same talker as used for the corresponding nonword in the old/new categorisation task in order to investigate whether the magnitude of lexical competition between existing and novel words was influenced by talker-specific details. If so, then this would suggest that lexical competition may be based on episodic representations rather than more abstract phonemic representations as is currently assumed by most models of the spoken word recognition. As noted above, Creel et al. (2008) have already demonstrated that talker-specific information can influence the amount of competition between two phonologically similar novel words in an eye-tracking paradigm. Experiment 2 extends this line of research by investigating whether talker-specific information can affect lexical competition between novel and existing words where the participant is required to consider the entire lexicon during spoken word recognition, rather than just the items presented on screen, as is the case in eye-tracking studies.

It was predicted that participants would show good recognition of the novel nonwords immediately after the exposure phase, as measured by accuracy in the old/new categorisation task, and that recognition rates would remain high over the course of the week, as has been observed in previous word learning studies (*e.g.*, Gaskell & Dumay, 2003). TSEs were also expected immediately after initial exposure, as in Experiment 1, with participants showing faster and more accurate categorisation of studied nonwords heard in the same voice at test. In terms of the

time-course of TSEs, if consolidation processes strengthen all aspects of novel lexical representations then TSEs may become stronger over time. Alternatively, if consolidation is a more selective process resulting in the establishment of more abstract representations, then TSEs may decline gradually over the course of a week, consistent with Goldinger's (1996) findings described in the introduction to this chapter. In contrast to TSEs, lexical competition effects for the basewords with novel nonwords competitors were expected to emerge only on Day 2, after a period of sleep-associated offline consolidation, consistent with previous studies (*e.g.*, Dumay & Gaskell, 2007). Talker-specific lexical competition would suggest that talker-specific details were retained within the neocortical system whilst talker-independent competition effects would suggest that representations within the neocortical system are more abstract in nature.

3.3.1 Method

Participants

Thirty-one undergraduate students (*age range* = 18–23 years, 9 male) from the University of York completed the experiment and were rewarded with either payment or partial course-credit. Six additional participants were tested but were removed from analyses due to failure to complete all three test session (5) or experimenter error (1).

Stimuli

The stimuli consisted of the 48 stimulus triplets used in Experiment 1. Due to unexpected significant main effects and interactions with list in the data from Experiment 1, the two lists of stimulus triplets were rearranged taking into account the size of the TSE for each item in the old/new categorisation task from Experiment 1. This TSE size was calculated by subtracting the mean categorisation RT when the test voice was the same as study from the mean RT when the test voice was different to study. As in Experiment 1, the two resultant lists of 24 stimulus triplets were matched on initial phoneme and number of syllables (12 bisyllabic and 12 trisyllabic items per list). The two lists were also matched as closely as possible in the number of phonemes (*M* = 7.96, *Range* = 6-11) and in frequency (*M* = 3.63, *Range* = 2-14) according to the CELEX database (Baayen, et al., 1993). Independent samples t-tests indicated that there were no significant differences between the two lists in either of

these variables (phonemes, t(46) = -.423, *ns*; frequency, t(46) = -1.799, *ns*) nor was there a difference between lists in the size of the TSE, t(45) = .633, *ns*, as calculated using the RT latencies from the old/new categorisation task in Experiment 1. However, the significant differences in acoustic duration (ms) of words spoken by the male talker and the female talker remained (basewords – t(47) = 6.227, $p < .001$; novel nonwords – t(47) = 3.735, $p < .001$; foil nonwords – t(47) = 4.172, $p < .001$), with items spoken by the male talker tending to be shorter than those spoken by the female talker.

Forty-eight monomorphemic English nouns were selected from the materials used by Tamminen and Gaskell (2008) as filler words for the lexical decision task (24 monosyllabic, 12 bisyllabic and 12 trisyllabic). When combined with the basewords from the stimulus triplets this resulted in 24 monosyllabic words, 36 bisyllabic, and 36 trisyllabic. Ninety-six nonword fillers that had created by changing either one or two syllables of existing words were also selected and were matched with the existing words in syllable length. An additional 30 fillers items (15 words and 15 nonwords) were used as practice items. All word and nonword fillers had been recorded by the male and female talkers from Experiment 1 during the same recording session as the stimulus triplets. As in Experiment 1, all audio files were digitized at a 44.1Hz sampling rate with 16-bit analogue-to-digital conversion, and peak amplitude was normalized using Adobe Audition.

Design

Each participant completed three sessions; one on Day 1, one on Day 2 approximately 24 hours later, and one on Day 8, one week after the first session. In the first session participants were familiarized with the novel nonwords in a phoneme monitoring task. Participants then complete the lexical decision task and old/new categorisation task immediately after study. On Days 2 and 8 participants completed only the lexical decision task and the old/new categorisation task. The order of these two tasks was fixed, with the lexical decision task always occurring before the old/new categorisation task in all three test sessions. Day 1 sessions lasted approximately 45 minutes, with Day 2 and Day 8 sessions taking around 15-20 minutes to complete.

Procedure

The *phoneme monitoring* task was identical to that used in Experiment 1, as was the manner in which changes in speaker between study and test were counterbalanced across participants. During the *test-phase* of the experiment the test voice was the same for all three items within a stimulus triplet. For example, if the novel word '*biscal*' was spoken in a male voice at test in the old/new categorisation task, then the foil nonword '*biscan*' was also heard in the male voice in the old/new categorization task, and the baseword '*biscuit*' was spoken in the male voice in the lexical decision task. Test-talker remained constant across all three test-session such that items classed as different-talker items in Session 1 remained in the opposite voice to study at all test-points, and same-talker items were heard in the same-voice as study at all time-points.

In the *lexical decision task* participants heard all 48 basewords, 48 word fillers, and 96 nonword fillers. Items were presented in a randomised order in two experimental blocks of 96 items that were matched in the number of test basewords, control basewords, word fillers, and nonword fillers. The order of the two experimental blocks was counterbalanced across participants. Half of the test basewords were heard in the male voice, and half in the female voice. The same was true for the control basewords, word fillers and nonword fillers, with the talker of each item counterbalanced across participants so that half of the participants heard each item in the male voice and half in the female voice. Note, the novel nonwords were *not* included in the lexical decision task.

The task began with a block of 30 practice trials to familiarise participants with the task. Participants were instructed to decide whether each item was an existing word or a made-up word, indicating their response by pressing the right or left button on the response pad respectively. RTs were recorded from word onset until a button-press response was made. The inter-trial interval was 500ms, with a maximum RT of 5s. Instructions emphasised both speed and accuracy. Feedback stating the mean RT and number of errors was provided after the practice block and at the end of each experimental block in order to encourage participants to maintain fast and accurate responding throughout the task.

The *old/new categorisation task* was identical to that used in Experiment 1 except that the delayed shadowing section of each trial was removed. As in

Experiment 1 participants were not informed about the manipulation of voices between study and test in order to avoid drawing attention to this variable.

3.3.2 Results

Study phase

Fifteen participants were exposed to List 1 (3 male) and 16 to List 2 (6 male). The mean error in the phoneme monitoring task was 5.6% (*SD* = 2.5%) indicating that participants were paying close attention to the phonological form of the novel nonwords. Error and RT data from the phoneme monitoring task were lost for one participant due to a technical failure at the end of the task. However, since the participants had completed the phoneme monitoring task at the time of technical failure data from this participant was still included in old/new categorization and lexical decision analyses.

A repeated-measures ANOVA, with factors study talker (male *vs.* female), and list (1 *vs.* 2), showed that there neither variable had an effect on accuracy in the phoneme monitoring task (talker – $F_1(1,28) = 3.51$, $p = .071$, $\eta_p^2 = .11$, $F_2(1,46) = 1.83$, *ns*; list – $F < 1$), indicating that any subsequent differences in performance in the test tasks are unlikely to be due to differences in encoding of male and female items. However, as in Experiment 1, RTs in the phoneme monitoring task differed significantly as a function of study talker (male: *M* = 1028ms, *SD* = 165ms; female: *M* = 1102ms, *SD* = 182ms), $F_1(1,28) = 110.84$, $p < .001$, $\eta_p^2 = .80$, $F_2(1,46) = 60.75$, $p < .001$, $\eta_p^2 = .57$, although, RTs did not vary between list, $F_1 < 1$, $F_2(1,46) = 3.15$, $p = .083$, $\eta_p^2 = .06$.

Talker-specificity effects

Data from the old/new categorisation task (in this and all subsequent experiments) were filtered using the same criteria as Experiment 1, and were analysed using SDT (Green & Swets, 1966). Overall participants responded correctly to 81.2% (*SD* = 7.9%) of the items. Three items were removed from the analysis as a result of having error scores more than 2.5 *SD* above the grand mean. Mean RT, measured from word onset, was 1205ms (*SD* = 201ms).

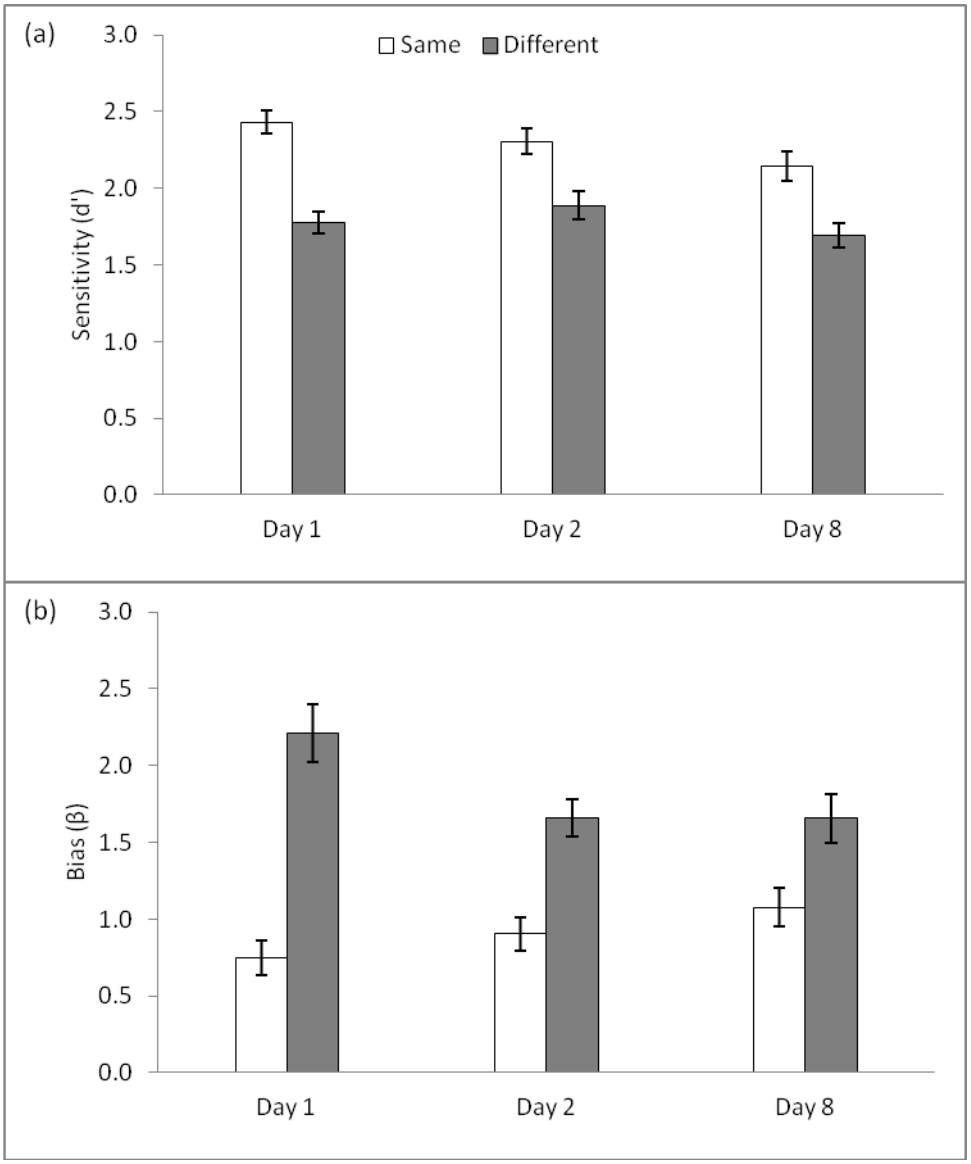

Figure 3.1. (a) Accuracy and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Both d′ and β data were analysed using a repeated measures ANOVA, with factors test-phase talker (same *vs.* different) and day (1, 2, or 8). For d′ data (Figure 3.1a) there was a significant main effect of test-phase talker, $F(1,29) = 28.45$, $p < .001$, $\eta_p^2 = .50$, with higher accuracy scores for same-talker items than different-talker items. There was also a marginal effect of day, $F(2,58) = 2.87$, $p = .064$, $\eta_p^2 = .09$, with d′ scored being significantly higher on Day 1 relative to Day 8, $F(1,29) = 4.23$, $p < .05$, $\eta_p^2 = .13$. All other comparisons between sessions were non-significant. However, there was no interaction between test-phase talker and day,

$F(2,58) = 1.31$, *ns*, suggesting that the same-talker advantage in d′ scores did not decrease across the week. Posthoc analyses confirmed this, revealing a significant main effect of test-phase talker at all time points (Day 1, $F(1,29) = 35.64$, $p < .001$, $\eta_p^2 = .55$, Day 2, $F(1,29) = 8.86$, $p < .01$, $\eta_p^2 = .23$; Day 8, $F(1,20) = 10.79$, $p < .01$, $\eta_p^2 = .27$).

For the β data (Figure 3.1b) there was a main effect of test-phase talker, $F(1,29) = 26.63$, $p < .001$, $\eta_p^2 = .48$, but a non-significant main effect of day, $F(2,58) = 1.10$, *ns.* However, the interaction between test-phase talker and day was significant, $F(1.6,47.6) = 7.12$, $p < .01$, $\eta_p^2 = .20$. Post-hoc analyses revealed the main effect of test-phase talker was significant on Day 1, $F(1,29) = 31.16$, $p < .001$, $\eta_p^2 = .52$, Day 2, $F(1,29) = 16.46$, $p < .001$, $\eta_p^2 = .36$, and on Day 8, $F(1,29) = 5.88$, $p < .05$, $\eta_p^2 = .17$, suggesting that whilst the difference between same and different talker items may decrease over the course of one week, the same-talker advantage was still robust on Day 8, supporting the results from the d′ analyses.

Lexical competition effects

In the lexical decision task participants performed accurately across all items, with a mean error score of 7.7% (*SD* = 4.0%). Only data from the 48 basewords were included in the lexical competition analysis, allowing comparison between words that had a novel competitor (*test basewords*) and words that did not have a novel competitor (*control basewords*). In this and all subsequent analyses of lexical competition data all incorrect responses were removed from the baseword data set prior to analysis, as were correct data points with a RT less than 200ms or more than 2.5 *SD* above or below the mean RT for each participant in each session. Finally, the lexical decision data were matched with the old/new categorisation responses; RTs to basewords corresponding to novel nonwords that participants did not correctly identify were removed from the data set on a session by session basis since increased lexical competition was not expected if participants did not recognise the novel nonwords. Overall 18.75% of data points were removed from the baseword data set.

A repeated measures ANOVA, with day (1, 2, and 8), and baseword type (test *vs.* control) included as within-participant variables revealed a significant main effect of day, $F_1(2, 58) = 13.57$, $p < .001$, $\eta_p^2 = .32$, $F_2(2,92) = 47.89$, $p < .001$, $\eta_p^2 = .51$. Further analysis revealed RTs were significantly slower on Day 1 compared to Day

2, $F_1(1,29) = 34.52$, $p < .001$, $\eta_p^2 = .54$, $F_2(1,46) = 91.69$, $p < .001$, $\eta_p^2 = .67$, and Day 8, $F_1(1, 29) = 8.75$, $p < .01$, $\eta_p^2 = .23$, $F_2(1,46) = 46.43$, $p < .001$, $\eta_p^2 = .51$. This effect may be due either to practice effects and task repetition resulting in decreased RTs on Days 2 and 8, or to fatigue on Day 1 due to participants having just completed the 20 min study phase of the experiment.

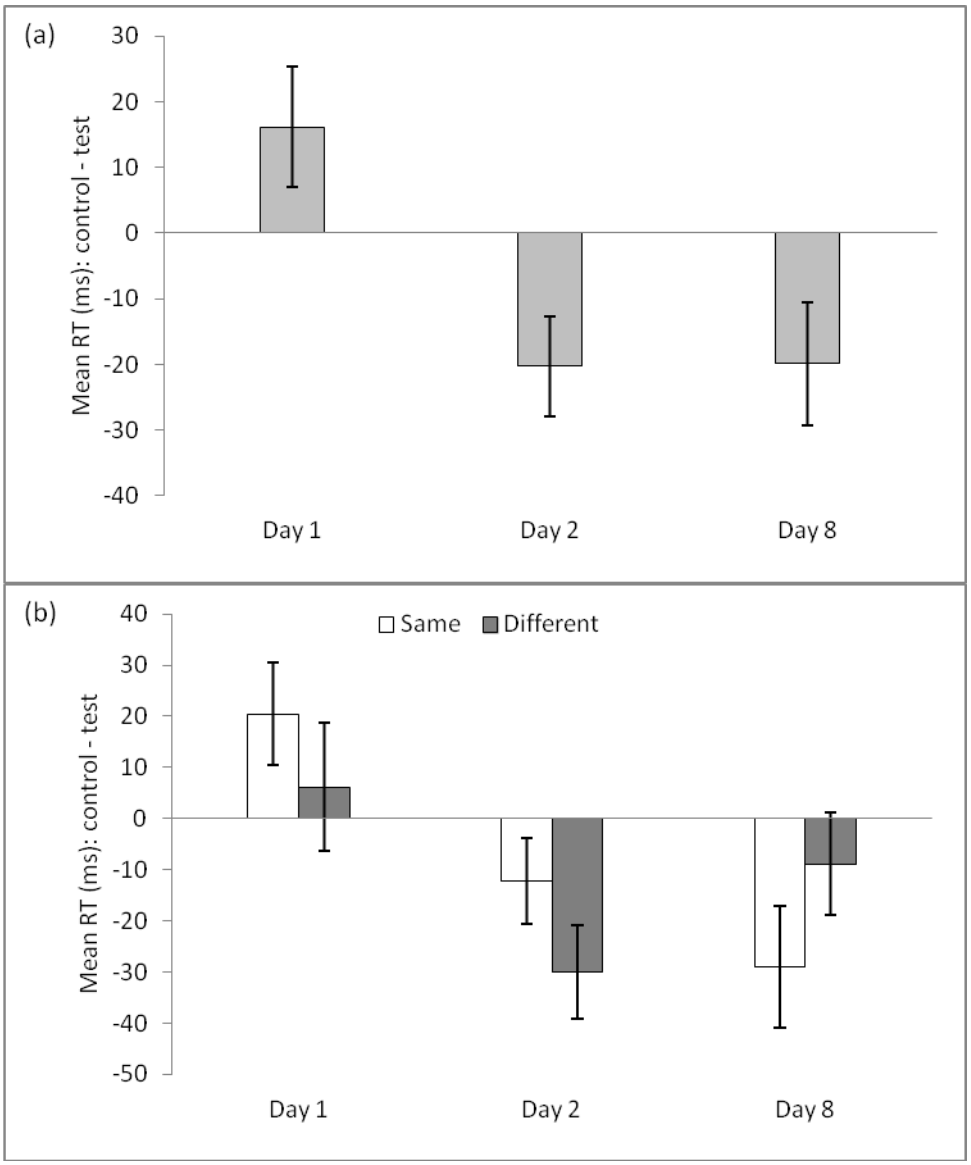

Figure 3.2. (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) basewords in the lexical decision task. (b) Lexical decision data split according to whether the test baseword was spoken in either the same voice that the corresponding novel word was trained in, or a different voice. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Although the main effect of word-type was not significant in the main analysis, the interaction between day and word-type was, $F_1(2, 58) = 9.95$, $p < .001$, $\eta_p^2 = .26$, $F_2(2,92) = 5.32$, $p < .01$, $\eta_p^2 = .10$, Further analysis revealed that the difference in RTs to test and control basewords was significant at all time points (Day 1, $F_1(1,29) = 4.96$, $p < .05$, $\eta_p^2 = .15$, $F_2(1,46) = 2.74$, *ns*; Day 2, $F_1(1,29) = 5.95$, $p < .05$, $\eta_p^2 = .17$, $F_2(1,46) = 6.75$, $p < .05$, $\eta_p^2 = .13$; Day 8, $F_1(1,29) = 6.26$, $p < .05$, $\eta_p^2 = .18$, $F_2(1,46) = 4.62$, $p < .05$, $\eta_p^2 = .09$). The interaction between word-type and day stems from the fact that RTs to test basewords were quicker than to control basewords on Day 1, but this pattern of responding reversed on Days 2 and 8, with faster RTs to control than test basewords at these two time points (Figure 3.2a). This pattern of data suggests that lexical competition between the novel nonwords and their phonologically similar basewords emerged only on Day 2, and remained constant across the remained of the week.

In order to determine whether talker-specific details influenced the size of the lexical competition effects observed, the test basewords were divided into two groups; those heard in the same voice that the corresponding novel nonword was heard in during study, and those heard in a different voice. A repeated measures ANOVA, with factors day (1, 2, 8), and baseword type (same-talker, different-talker, and control) showed that the main effect of baseword type was not significant, $F < 1$, but that there was a significant interaction between baseword type and day, $F_1(2.7,78.4) = 3.06$, $p < .05$, $\eta_p^2 = .10$, $F_2(4,180) = 3.46$, $p < .01$, $\eta_p^2 = .07$ (Figure 3.2b). In order to explore this interaction further the data from each test session were analysed separately for same-talker and different-talker basewords. Priming or facilitatory effects on Day 1 approached significant only for same-talker items, $F_1(1,29) = 3.45$, $p = .074$, $\eta_p^2 = .11$, $F_2(1,46) = 3.80$, $p = .057$, $\eta_p^2 = .08$. Lexical competition was significant only for different-talker items on Day 2, $F_1(1,29) = 7.32$, $p < .05$, $\eta_p^2 = .20$, $F_2(1,46) = 4.54$, $p < .05$, $\eta_p^2 = .09$, but only same-talker items on Day 8, $F_1(1,29) = 4.93$, $p < .05$, $\eta_p^2 = .15$, $F_2(1,46) = 6.29$, $p < .05$, $\eta_p^2 = .12$.

Correlations between talker-specificity and lexical competition effects

In order to examine whether changes in TSEs (particularly β, where there was a significant interaction between day and word-type) correlated with changes in lexical competition across the course of the week correlational analyses were

conducted. For the lexical competition data, differences between the size of the lexical competition effect were calculated for Day 1-Day 2, Day 1-Day 8, and Day 2-Day 8. For the old/new categorisation data, a difference score was calculated for each day by subtracting the d′, β, or novel nonword RT score in the different-talker condition from the scores in the same-talker condition. Differences across day were then calculated as for the lexical competition measures. No significant correlations were observed, suggesting that these two effects are follow independent time-courses.

3.3.3 Discussion

As in Experiment 1 the data from Experiment 2 show that accurate phonological representations are formed immediately upon encountering novel words, with talker information being encoded and stored alongside this phonological information. Building on this finding, Experiment 2 demonstrates that talker-specific information can affect recognition of newly-learned words up to one week later, as demonstrated by the significant same-talker advantage observed in both the speed and accuracy with which the new words were categorized in all three test sessions (for RT data see Appendix C). The finding that TSEs for novel nonwords were significant even one week after initial exposure to the items stands in contrast to data from Goldinger's (1996) old/new categorisation task, described in the introduction to this chapter, showing that TSEs for existing words declined over the course of a week in a similar old/new categorisation task. This discrepancy will be discussed further in the introduction to Chapter 4, with Experiments 4 and 5 providing a direct comparison of TSEs for existing and novel words in order to determine whether the different time-courses observed are due simply to methodological differences, or whether they are driven by the nature of the words themselves. That is, the difference in time-course of TSEs for existing and novel words may be due to the fact that recognition of existing words is likely to rely primarily on pre-established neocortical representation whereas recognition of novel words is likely to rely to a greater extent on hippocampal-mediated representations.

In comparison to the immediacy of TSEs for the novel nonwords, lexical competition did not emerge until Day 2. This finding is consistent with previous research indicating that a period of sleep-associated offline consolidation is required

in order for novel nonwords to become integrated within the existing lexicon, and to begin interacting with phonologically-similar words during spoken word recognition (*e.g.*, Dumay & Gaskell, 2007). One point to note is that there was a significant facilitatory effect for test basewords on Day 1, with RTs being quicker to test than control basewords. The most plausible explanation for this finding is that participants may have become aware of the similarity between the novel nonword (*e.g., biscal*) and their phonologically-similar basewords (*e.g., biscuit*) during the study phase of the experiment. Even if participants were not consciously aware of this similarity, hearing the novel items repeatedly 18 times during study is likely to have partially activated, or primed the phonologically-similar test basewords such that they were activated more rapidly than the control basewords in the subsequent lexical decision task on Day 1.

Interestingly, the lexical competition effects observed on Days 2 and 8 differed depending on whether the test baseword was heard in either the same or a different voice to that in which the phonologically-similar novel nonword was studied; only different-talker items engaged in lexical competition on Day 2, but conversely only same-talker items engaged in competition on Day 8. This pattern of data is rather difficult to interpret, and will be discussed in more detail in Chapter 5 following cross-experiment analyses of all of our lexical decision data.

The emergence of talker-specific lexical competition in Experiment 2 builds on Creel et al.'s (2008) eye-tracking study, which showed talker-specific lexical competition only between pairs of phonologically similar novel words. In comparison, our experiment compared lexical competition between existing and novel item. Moreover, in Creel et al.'s study participants were required to consider only a closed set of lexical items during recognition of the novel spoken words. Data from Experiment 2 suggest that even when the whole lexicon must be considered, competition processes are still talker-specific even after a period of sleep-associated offline consolidation. However, it is possible that using two talkers in the lexical decision task slowed processing in Experiment 2. Martin et al. (1989) claim that more processing resources in working memory are required when a list of words is spoken by multiple talkers compared to only a single talker. As such, processing of items within multiple-talker lists is likely to be slower and more effortful. Consistent with this suggestion, RTs in our lexical decision task were already approximately 150ms slower than previous studies that have used lexical decision to examine the

emergence of lexical competition using only a single talker (*e.g.*, Gaskell & Dumay, 2003). According to Luce et al.'s (2003; McLennan & Luce, 2005) time-course hypothesis, this would have allowed extra time for talker-specific details to be integrated with the retrieved phonological representations, enabling TSEs in lexical competition measures to be observed. Experiment 3 addressed this possibility.

3.4 Experiment 3

In this experiment changes were made to the lexical competition test in order to address the fact that the RTs observed in Experiment 2 were much longer than found in previous studies using similar stimuli (*e.g.*, Gaskell & Dumay, 2003). Experiment 3 also aimed to investigate the contributions of recollection and familiarity to recognition memory for novel words, both immediately after exposure, as well as one week later.

With regards to our first aim, two changes were made to the lexical competition task. Firstly, only one talker was used throughout the task, with half of the participants hearing only the male talker during this task, and the other half hearing only the female talker. It is important to note that the end result of this manipulation of talker is the same as in Experiment 2 with half of the basewords heard in the same voice as the studied novel nonwords, and half heard in a different voice. As such, it was still be possible to examine whether lexical competition between existing and novel words was talker-specific.

In addition to this change, a different measure of lexical competition, pause detection (Mattys & Clark, 2002), was used in order to determine whether the intriguing and somewhat unexpected pattern of talker-specific lexical competition effects observed in Experiment 2 could be replicated using a different task. In the pause detection task participants must monitor for short 200ms pauses artificially embedded within spoken words. Mattys and Clark (2002) found that words with late uniqueness points showed longer pause detection latencies that words with early uniqueness points. They suggested that pause detection latencies are longer in cases where the active lexical candidate set is larger because processing of multiple candidates uses up processing resources that would otherwise be allocated to pause detection. Word learning studies in adults have assumed that by adding a novel competitor to the lexicon, the number of lexical candidates for the phonologically-similar baseword increases. As a result, additional resources are required to process

the existing baseword, leaving fewer resources for pause detection, thus resulting in increased RTs once the novel nonword has been integrated into the existing lexicon (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Henderson et al., submitted-b).

Pause detection has a number of advantages over lexical decision as a measure of lexical competition. Firstly, pause detection provides an online measure of the amount of lexical activity at the time-point in the speech signal at which the pause is inserted, whereas lexical decision is only able to provide a measure of the amount of lexical competition at the end point of spoken word recognition. As such, if talker information is integrated with the retrieved representation at a late point in spoken word recognition, as predicted by Luce et al.'s (2003; McLennan & Luce, 2005) time-course hypothesis, then talker-specific lexical competition should not be observed in pause detection. Secondly, pause detection does not require a meta-linguistic judgment, unlike lexical decision, which requires participants to make a decision (word *vs.* nonword) that explicitly taps lexical processing, with different processes required to make '*word*' and '*nonword*' decisions (Marslen-Wilson & Warren, 1994). That is, '*word'* responses can only be made once the initial cohort of lexical candidates has been narrowed down, and a single lexical item has been selected as the target word. '*Nonword*' responses, on the other hand require listeners to determine that none of the initial cohort of items match the speech input. Pause detection does not require an explicit judgment to be made about the linguistic or lexical properties of the speech input.

In addition to the changes made to the lexical competition task, Experiment 3 also investigated the contributions of recollection and familiarity to recognition memory for novel words. A number of researchers have drawn a distinction between recollection and familiarity within the context of recognition tasks such as the old/new categorisation task (Brandt, Gardiner, Vargha-Khadem, Baddeley, & Mishkin, 2009; Daselaar, Fleck, & Cabeza, 2006; Diana, Yonelinas, & Ranganath, 2007; Eichenbaum, Yonelinas, & Ranganath, 2007; Elfman, Parks, & Yonelinas, 2008; Ranganath et al., 2004; Yonelinas, 2002; Yonelinas, Otten, Shaw, & Rugg, 2005). Items are said to be recollected only if specific contextual details, such as talker identity, are recalled. Familiarity, on the other hand, corresponds to a non-specific sense that an item has been recently encountered or, in other words, recognition without recovery of any specific contextual details from the encoding

episode. One way of investigating whether an item has been recollected or has been recognised simply on the basis of familiarity is to ask participants to rate their confidence in the old/new categorisation response and to plot receiver operating characteristic (ROC) curves from this data (Yonelinas, 2002). Recollected items are expected to have a high confidence rating, whereas items recognised on the basis of familiarity will show lower confidence ratings. Moreover, confidence ratings for recollected items should always be high, resulting in a non-linear confidence rating function, whereas confidence ratings for familiar items should be spread over a range of confidence intervals, resulting in a more linear confidence rating function (Diana et al., 2007). Another task that can be used to investigate the contribution of recollection to recognition memory is to ask participants source memory questions (*i.e.*, to recall specific information about the previous encounter with an item). Where an item is identified on the basis of familiarity participants should be unable to make a source memory response (Diana et al., 2007).

It has been proposed that recollection and familiarity engage distinct brain mechanisms (Yonelinas et al., 2005). Recollection of source memory (or 'episodic') details is thought to rely heavily on the hippocampus (Davachi, Mitchell, & Wagner, 2003; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Ranganath et al., 2004), and to be supported by the sparse, pattern-separated representations formed by the hippocampus when information is initially encoded (Bakker et al., 2008; Elfman et al., 2008). In support of this suggestion patients with developmental amnesia, in which there is selective damage to the hippocampus at birth or during early childhood, typically show better recognition than recall (Adlam, Malloy, Mishkin, & Vargha-Khadem, 2009). Recall is thought to be more dependent on the precise recollection of an item where as recognition relies to a greater extent on familiarity processes that are assumed to be more dependent on the surrounding medial temporal lobe (MTL)[3] (Brandt et al., 2009; Elfman et al., 2008; Henson, Cansino, Herron, Robb, & Rugg, 2003; Ranganath et al., 2004), particularly the perirhinal cortex (Aggleton et al., 2005; Davachi et al., 2003). Nevertheless, one potential problem with studying patients who suffered hippocampal damage at an early age is that the neural plasticity of developing brains may have allowed for

[3] McClelland, et al. (1995, p. 423) note that whilst the perirhinal and parahippocampal cortices are anatomically defined as neocortex, it may be best to consider these two regions as areas in which the hippocampus and neocortical networks overlap.

some functional reorganization of cognitive abilities in order to compensate for the damage incurred (see Maguire, Vargha-Khadem, & Mishkin, 2001; Manns & Squire, 1999). However, hippocampal lesions acquired during adulthood have also been found to result in impaired recollection but preserved familiarity (Aggleton et al., 2005; Bastin et al., 2004; Holdstock, Mayes, Gong, Roberts, & Kapur, 2005; Mayes, Holdstock, Isaac, Hunkin, & Roberts, 2002; Mayes et al., 2004; Yonelinas et al., 2005), as have selective hippocampal lesions in rats (Fortin, Wright, & Eichenbaum, 2004), supporting the conclusions drawn from research looking at developmental amnesia (but see Manns & Squire, 1999, for evidence that hippocampal damage can result in both familiarity and recollection deficits). In comparison, Bowles et al. (2007) demonstrated that a patient with selective preservation of the hippocampus following resection of anterior medial-temporal lobe showed preserved recollection, but impaired familiarity scores. Additional evidence supporting this dissociation comes from a functional imaging study by Davachi and colleagues (2003) showing that encoding activation in the hippocampus and posterior hippocampal cortex predicted later source recollection but not word recognition whilst encoding activity in the perirhinal cortex predicted the reverse. Taken together these findings demonstrate a double dissociation between recollection and familiarity being mediated by hippocampal and neocortical/surrounding MTL regions respectively, a dissociation that fits nicely with a hybrid CLS model of lexical representation in which both episodic and abstract memory representations may co-exist.

In order to investigate the contributions of recollection and familiarity to recognition memory for novel words participants were required to make three responses to each item presented in the old/new categorisation task in Experiment 3. First, participants judged the item as old or new, as in Experiments 1 and 2. Subsequently, participants rated how confident they were about the accuracy of their old/new judgment on a 1 to 7 scale, with 1 corresponding to '*definitely new*' and 7 corresponding to '*definitely old*'. Finally, participants were also asked to indicate whether the item had been heard in a male or a female voice during the phoneme monitoring task. This source memory task was included in order to investigate whether talker information from the study phase of the experiment could be explicitly recollected. Whilst Experiment 2 clearly demonstrated that participants encoded and stored detailed talker-specific information when exposed to novel words, and that this information was able to influence recognition of the novel words

up to one week later, it is not clear whether participants were able to explicitly access this talker information at all time points.

It was predicted that participants would show both accurate recognition of the novel nonwords and TSEs for these novel words immediately after study, consistent with Experiment 1, and that TSEs would still be observed one week later in the old/new categorisation task, consistent with Experiment 2. However, if recollection processes are indeed dependent on the hippocampus, and familiarity processes by cortical or MTL regions, then we would expect that the number of correct male/female responses to decrease over the course of one week, and that confidence ratings would becoming less certain, and thus move towards the mid-point of our seven point confidence rating scale, indicating that recognition of the novel words was less reliant on hippocampally-mediated representations one week post-exposure to the novel words.

Lexical competition effects on the other hand were expected to be absent immediately after exposure to the novel nonwords, but to emerge on Day 2 after a period of sleep-associated offline consolidation. Based on the findings from Experiment 2, it was predicted that only different-talker items would engage in competition on Day 2, and only same-talker items would engage in competition on Day 8.

3.4.1 Method

Participants

Forty-eight adults (*age range* = 18-25 years, 15 male) recruited from the University of York and surrounding areas completed the experiment and were rewarded with either payment or partial course credit. Nineteen additional participants were tested. Data from 13 were lost due to equipment failure, and the remaining 6 participants were replaced due to failure to complete all three test sessions (2), experimenter error (2), or scoring more than 2.5 *SD* above the mean error score in the phoneme monitoring task (2), which suggested that the novel nonwords had not been correctly encoded during the study task.

Stimuli

All stimuli and counterbalancing of stimuli and talker for study and test was identical to Experiment 2.

Design and Procedure

The design was the same as Experiment 2 except that pause detection replaced lexical decision, and confidence rating and male/female categorization tasks were included after each old/new decision.

In the *pause detection* task participants heard all 48 basewords alongside 68 filler items. All fillers were existing words, and comprised the 48 filler words and 20 practice words from the lexical decision task in Experiment 2. Twenty additional words were recorded as practice items by the two talkers from Experiments 1 and 2. For the basewords, 200ms pauses were inserted just before the final vowel of the bisyllabic items (N = 24), and either after the second syllable (N = 12) or just before the final vowel (N = 12) of the trisyllabic items. For the filler items pauses were inserted either before (N = 9) or after (N = 20) the vowel of monosyllabic words, before the first vowel (N = 10) or after the last vowel (N = 10) of bisyllabic words, and after the first syllable (N = 10), second syllable (N = 5), or before the final vowel (N = 4) of the trisyllabic words. Items were presented in a randomised order in two experimental blocks of 58 items that were matched in the number of test basewords, control base-words, and fillers. The order of the two experimental blocks was counterbalanced across participants. Each word was heard only once. Half of the participants heard each baseword with a pause inserted in it, and half heard the baseword with no pause. Additionally, half of the participants heard all items spoken by the male talker, and the other half heard only the female speaker.

The task began with a block of 20 practice trials to familiarise participants with the task. Participants were instructed to decide whether each item contained a short 200ms pause or not, indicating their response by pressing the right or left button on the response pad respectively. As in Gaskell and Dumay's (2003) study participants were required to respond on both pause-present and pause-absent trials, since "...if overall lexical activity makes use of shared resources that are also required for detecting pauses, then a similar delay should be observed when participants are required to respond in pause-absent trials. A further advantage of requiring a response in pause-absent trials is that behaviour can be observed in cases where the basewords are presented whole" (Gaskell & Dumay, 2003, p.119). RTs were recorded from pause onset, as indicated by a digital marker inserted in the audio file, until a button-press response was made. The inter-trial interval was 1s, with a maximum RT of 3s. Instructions emphasised both speed and accuracy. Feedback

stating the mean RT and number of errors was provided after the practice block and at the end of each experimental block to encourage participants to maintain fast and accurate responding throughout the task.

In the *old/new categorisation* task participants were asked to make three responses to each item heard. Each trial began with a central fixation cross (+), indicating that a word was about to be presented. After hearing the word participants classified the word as '*old*' or '*new*', with RTs measured from word-onset until a button-press response was made. Participants then rated the confidence in their old/new judgement using a scale from 1 to 7, with 1 corresponding to '*definitely new*', and 7 to '*definitely old*'. Participants were instructed to try to use the full range of responses offered by the rating scale. Finally, participants indicated whether they thought that the word had previously been heard in a male or a female voice during the phoneme monitoring task. Instructions stated that participants should ignore the voice used in the old/new categorization task, instead focusing on which talker they thought had spoken the word during the phoneme monitoring task only. Thus, contrary to Experiments 1 and 2, the manipulation of voice was made clear to participants in this experiment. Participants were instructed to make a male/female judgment for all items since it may be that participants were able to correctly identify the talker that the word was originally heard in whilst incorrectly classifying the words as '*new*', or vice versa. For confidence rating and male/female judgments RTs were measured from the appearance of an onscreen cue (either the 1-7 scale, or the words '*male*' and '*female*') until a button-press response was made. All responses in this task were made using labelled keys on the computer keyboard. Each nonword was presented only once at the beginning of each trial, before the first decision, the old/new categorization, was made. The interval between each section of a trial was 500ms, as was the inter-trial interval. The maximum RT for each section of a trial was 5s. Instructions emphasised that participants should respond as quickly and accurately as possible for the first two decisions, but to focus on accuracy when making a decision about the study voice of the item.

3.4.2 Results

Study phase

Twenty-four participants were exposed to List 1 (7 male) and 24 to List 2 (8 male). The mean error rate in the phoneme monitoring task was 5.3% (*SD* = 2.8%),

indicating that participants were paying close attention to the phonological form of the novel nonwords during the study phase of the experiment. A repeated-measures ANOVA, with factors study talker (male *vs.* female), and list (1 *vs.* 2) showed that there was main effect of study-phase talker in the by-participants analysis, $F_1(1,46) = 9.37$, $p < .01$, $\eta_p^2 = .17$, with more errors being made to items heard in the female voice (*M* = 6.0%, *SD* = 3.5%) than items heard in the male voice (*M* = 4.7%, *SD* = 2.6%) at study. Likewise, the main effect of list was significant in the by-participants analysis, $F_1(1,46) = 6.28$, $p < .05$, $\eta_p^2 = .12$, with more errors being made by participants hearing List 2 (*M* = 6.2%, *SD* = 3.4%) than participants hearing List 1 (*M* = 4.4%, *SD* = 2.5%). However, neither of these main effects were significant in the by-items analysis (talker - $F_2(1,46) = 2.35$, *ns*; list - $F_2 < 1$), and the interaction between study talker and list was not significant in either analysis, $F_1(1,46) = 2.71$, *ns*, $F_2(1,46) = 2.49$, *ns*.

The mean RT in the phoneme monitoring task was 1139ms (*SD* = 254ms). As in Experiments 1 and 2, analysis of RTs revealed a main effect of study talker, $F_1(1,46) = 88.07$, $p < .001$, $\eta_p^2 = .66$, $F_2(1,46) = 7.38$, $p < .01$, $\eta_p^2 = .14$, with RTs being faster to male items (*M* = 1109ms, *SD* = 253ms) than female items (*M* = 1169ms, *SD* = 257ms). The main effect of list was significant only in the by-items analysis, $F_2(1,46) = 24.61$, $p < .001$, $\eta_p^2 = .35$, with participants responding faster to List 1 items (*M* = 1096ms, *SD* = 244ms) than to List 2 items (*M* = 1182ms, *SD* = 262ms). However, the interaction between study talker and list was not significant, $F < 1$, indicating that any advantage seen for items heard in the male voice at study was similar for both lists of items.

Talker-specificity effects

For each word heard in the old/new categorization task participants made three responses; an old/new judgment, a confidence rating corresponding to the old/new judgment, and a male/female categorisation decision. Data from each of these responses were analysed separately.

In the *old/new categorization* task three items produced error scores more than 2.5*SD* above the grand mean. Data from these three items were removed prior to analysis. After removal of these items participants responded correctly to 83.1% (*SD*

= 6.0%) of the items. RTs were measured from word onset until participants made an old/new categorisation button-press response (*M* = 1820ms, *SD* = 300ms)[4].

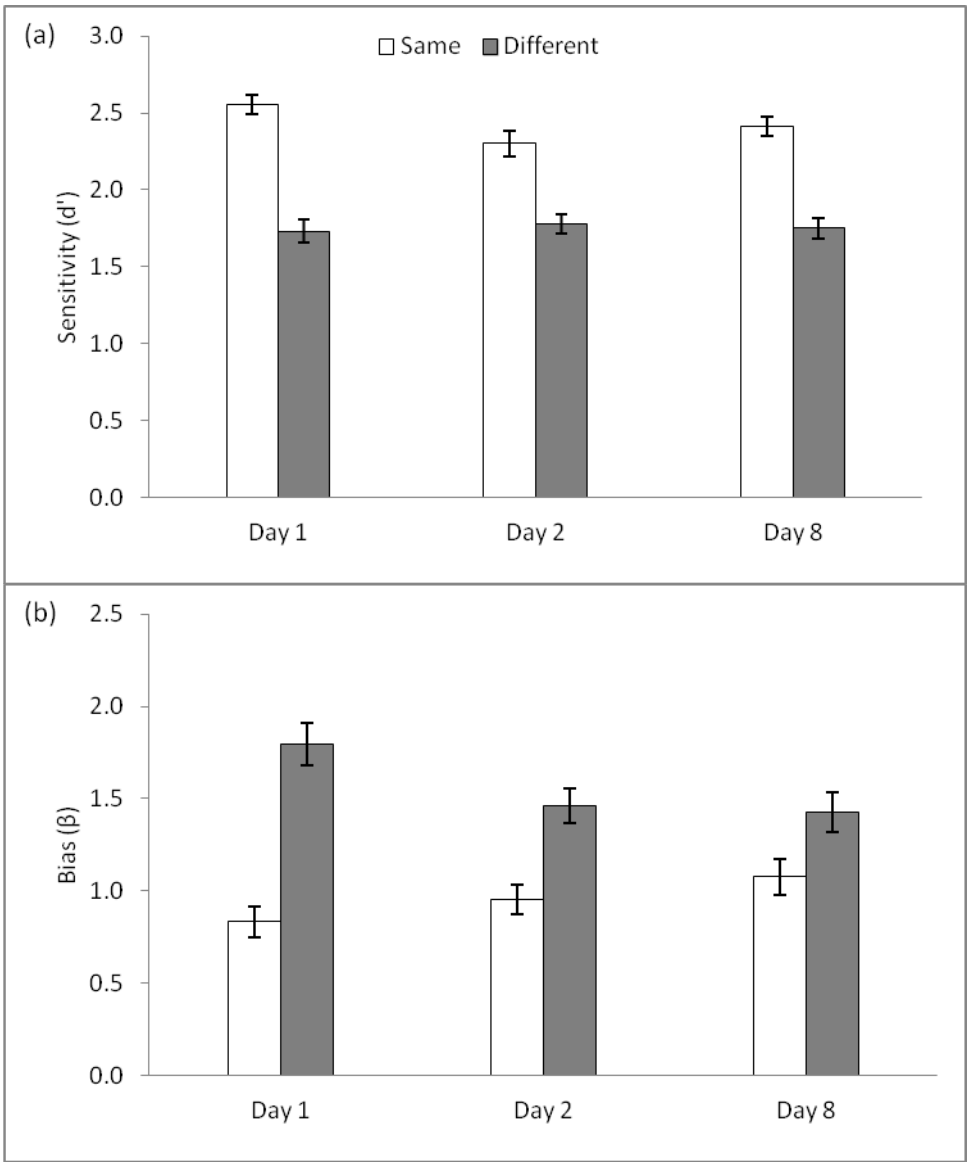

Figure 3.3. (a) Accuracy and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Analysis of d′ values (Figure 3.3a) revealed a significant main effect of test-phase talker, $F(1,46) = 66.63$, $p < .001$, $\eta_p^2 = .59$, but no main effect of day, $F(2,92)$

[4] RTs in the old/new categorisation task were much longer than those observed in Experiments 1 and 2 where only the old/new categorisation decision was required. It seems likely that requiring participants to make three decisions to every word heard (old/new, confidence rating, and male/female) places increasing demands on processing, thus slowing down responses in all tasks.

= 1.07, *ns*, and no interaction between test-phase talker and day, F(2,92) = 2.42, *ns*. These findings indicate that overall accuracy did not change over the course of the week, that participants were better at correctly classifying same-talker items as either old or new than different-talker items, and that this same-talker advantage did not change over the course of a week. All of these findings are consistent with Experiment 2 despite the fact that participants were explicitly instructed to ignore information about the study talker in the test phase of Experiment 3, indicating that TSEs are maintained over time for novel nonwords regardless of whether participants are instructed to ignore talker information at test or not.

For the β values there was a significant main effect of test-phase talker, F(1,46) = 31.52, p < .001, η_p^2 = .41, and a non-significant main effect of day, F(2,92) = .58, *ns*. However, the interaction between test-phase talker and day was significant, F(2,92) = 4.97, p <.05, η_p^2 = .10. Further analysis revealed that main effect of test-phase talker was significant on Days 1, F(1,46) = 38.27, p < .001, η_p^2 = .46 and 2, F(1,46) = 13.29, p = .001, η_p^2 = .22, but was only marginally significant on Day 8, F(1,46) = 3.87, p = .055, η_p^2 = .08), suggesting that there was a change in bias across the course of the week, with participants gradually adopting more similar biases when responding to same- and different-talker items as the week progressed (Figure 3.3b). This is somewhat different to Experiment 2 where the main effect of test-phase talker was significant at all time points. Nonetheless, even in Experiment 2 the TSE effect size decreased numerically across the course of the week and the interaction between day and test-phase talker was significant.

Confidence ratings were used to generate ROC curves. The hit rate (*'true positive rate'*) was plotted as a function of the false-alarm rate (*'false positive rate*') for each of the seven points on the confidence rating scale. In order to analyse the data the area under the ROC curve (AUC) was calculated in SPSS, which uses a bi-negative exponential model to fit a curve to the data, for each participant individually, with the data split by test-phase talker (same-talker and different-talker) and day (1, 2, and 8). Mean AUC values for each condition are plotted in Figure 3.4. AUC values provide an additional measure of sensitivity that is unaffected by response bias. Moreover, higher AUC scores reflect better recognition performance, with high AUC scores often taken as indicating a greater contribution of recollection to recognition memory and low AUC scores reflecting a greater reliance on

familiarity. Note that values of 0.5 reflect chance performance in the analysis of AUC scores, and thus represent an inability to discriminate between studied and unstudied items.

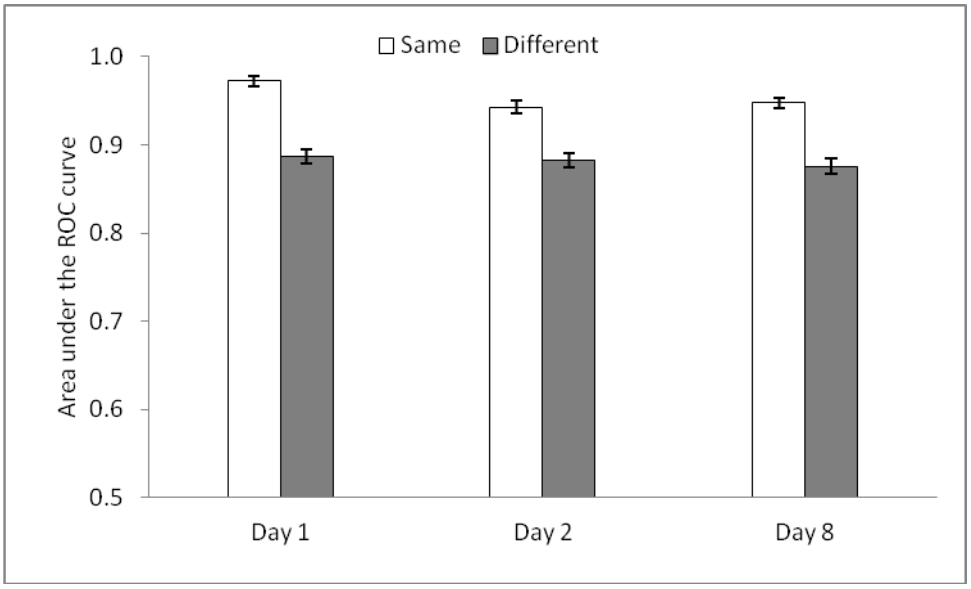

Figure 3.4. Confidence ratings given to old/new categorisation decisions were used to plot ROC curves. Area under the ROC curve was calculated separately for same and different talker items in each test session. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

A repeated-measures ANOVA, in which area under the curve (AUC) values were included as the dependent variable and factors day (1, 2, 8), and test-phase talker (same *vs.* different) as within-participant variables revealed that there was a significant main effect of test-phase talker, $F(1,45) = 82.53$, $p < .001$, $\eta_p^2 = .65$, with greater AUC values for same-talker items, indicating a greater contribution of recollection processes to recognition of these items. The main effect of day was also marginally significant, $F(2,90) = 2.89$, $p = .06$, $\eta_p^2 = .60$, with further analysis revealing a significant decrease in AUC scores between Days 1 and 2, $F(1,45) = 4.08$, $p < .05$, $\eta_p^2 = .08$, but no further change from Day 2 to Day 8, $F < 1$. These findings suggest that participants become less reliant on recollection processes, and more reliant on familiarity-based mechanisms between Days 1 and 2. Nevertheless, the interaction between test-phase talker and day was not significant, $F(2,90) = 1.63$, *ns*, indicating that the size of the same-talker advantage itself did not change over the

course of a week, suggesting that talker-information was still important in recognition of previously-studied words at all time-points.

In the *male/female categorization* analysis, only data from novel nonwords were included since the foil nonwords were encountered for the first time in the categorization task on Day 1. Mean accuracy was 66.4% (*SD* = 12.4%), significantly above chance, $t(49) = -9.29$, $p < .001$. Since participants were instructed to focus on accuracy rather than speed when making a male/female decision RT latencies were not analysed.

Accuracy scores, calculated separately for same- and different-talker items in each test session are reported in Table 3.3. A repeated-measures ANOVA with variables test-phase talker (same *vs.* different) and day (1, 2, 8) revealed that there was a significant main effect of test-phase talker, $F_1(1,46) = 78.72$, $p < .001$, $\eta_p^2 = .63$, $F_2(1,46) = 167.26$, $p < .001$, $\eta_p^2 = .78$; unsurprisingly participants responding more accurately to items heard in the same voice as study (*M* = 80.8%, *SD* = 15.7%) compared to items heard in a different voice to study (*M* = 52.6%, *SD* = 22.0%). There was also a significant main effect of day, $F_1(1.6, 73.1) = 11.01$, $p < .001$, $\eta_p^2 = .19$, $F_2(2,92) = 7.97$, $p < .001$, $\eta_p^2 = .15$. Posthoc comparisons revealed significant, or marginally significant, decreases in accuracy between both Days 1 and 2, $F_1(1,46) = 3.94$, $p = .053$, $\eta_p^2 = .08$, $F_2(1,46) = 5.98$, $p < .05$, $\eta_p^2 = .12$, and Days 2 and 8, $F_1(1,46) = 8.41$, $p < .01$, $\eta_p^2 = .16$, $F_2(1,46) = 3.15$, $p = .083$, $\eta_p^2 = .06$, indicating that participants got progressively worse at this task as the week progressed. Nevertheless, the interaction between test-phase talker and day was not significant, $F < 1$, suggesting that the same-talker advantage did not decrease over the course of one week as overall accuracy in the task decreased.

Table 3.3. Percentage of correct responses in the male/female categorisation task (novel nonwords only), split according to whether the item was spoken in the same or a different talker to study.

Day	Same	Different
1	83.5	57.0
2	81.1	52.1
8	77.6	48.6

Lexical competition effects

In the pause detection task participants performed accurately across all items, with a mean error score of 7.4% (*SD* = 7.4%). One participant had an error score of approximately 50% in all three test sessions; data from this participant was removed from further analyses. Only data from the 48 basewords were included in the lexical competition analysis, as in analysis of lexical decision data in Experiment 2. Both pause present and pause absent items were included in the analysis. Incorrect responses and correct data points with an RT more than 2.5 *SD* above or below the mean RT for each participant in each session were removed (6.29% of the data points).

For remaining data points RTs were calculated from pause onset (*cf.* Gaskell & Dumay, 2003), and all negative RTs, indicating that participants responded before they could be certain whether or not there was a pause in the item, were removed. Two participants were removed at this point for having an overall error (1) or RT (1) score more than 2.5 *SD* above the grand mean. Finally, the pause detection data were matched with the old/new categorisation responses; RTs to basewords corresponding to novel nonwords that participants did not correctly identify were removed from the data set on a session by session basis since increased lexical competition was not expected for these items. Overall 14.7% of the data points were removed from the baseword data set.

A repeated-measures ANOVA, with factors day (1, 2, 8) and baseword type (test *vs.* control) revealed that there was significant main effect of day, $F_1(1.4,62.2) = 7.19$, $p = .001$, $\eta_p^2 = .14$, $F_2(2,90) = 61.75$, $p < .001$, $\eta_p^2 = .58$. Additional analysis revealed that response times on Day 1 differed significantly from Day 2, $F_1(1,43) = 9.41$, $p < .01$, $\eta_p^2 = .18$, $F_2(1,45) = 94.11$, $p < .001$, $\eta_p^2 = .68$, and Day 8, $F_1(1, 43) = 6.77$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,45) = 70.79$, $p < .001$, $\eta_p^2 = .61$, with overall RTs being slower on Day 1, as in Experiment 2.

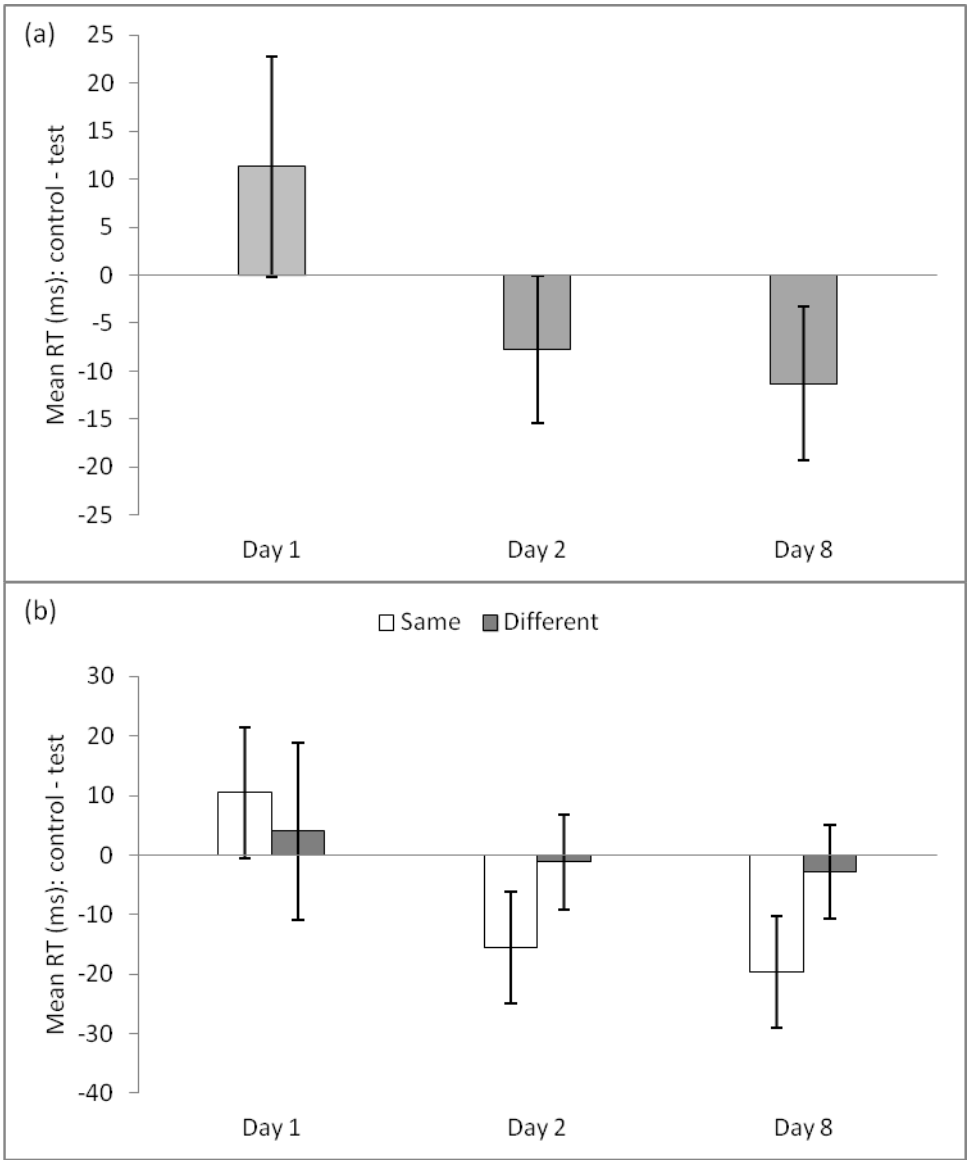

Figure 3.5. (a) Mean difference between response times to control (no novel competitor) and test (novel competitor) base-words in the pause detection task. (b) Pause detection data split according to whether the test baseword was spoken in the same voice that the corresponding novel word was trained in, or a different voice. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

The main effect of baseword type was not significant, $F < 1$, although the interaction between day and word-type was marginally significant in the by-participants (but not by-items) analysis, $F_1(1.7,74.8) = 3.06$, $p = .052$, $\eta_p^2 = .07$, $F_2 < 1$, indicating that RTs to test basewords were quicker than to control basewords on Day 1, but this pattern of responding reversed on Days 2 and 8, with faster RTs to control than test basewords at these two time-points (see Figure 3.5a). However,

comparison of RTs to test and control basewords at each time-point individually did not reveal any significant effects (Day 1, $F_1(1,43) = 1.71$, *ns*, $F_2(1,45) = 1.04$, *ns*; Day 2, $F_1(1,43) = 1.13$, *ns*, $F_2 < 1$; Day 8, $F_1(1,43) = 1.71$, *ns*, $F_2 < 1$), suggesting that robust lexical competition effects were not observed at any time point in the pause detection task.

As in Experiment 2, additional analyses were conducted in order to determine whether RTs in the pause detection task were influenced by whether the baseword was heard in a same or different talker to that in which the novel nonword was studied (Figure 3.5b). The main effect of baseword type (same-talker, different-talker, control) was not significant, $F_1(2,86) = 1.28$, *ns*, $F_2 < 1$, nor was the interaction between baseword type and day, $F_1(3.1,131.8) = 1.30$, *ns*, $F_2 < 1$, indicating that no lexical competition was observed when the test basewords were split into same- and different-talker items, unlike Experiment 2.

Correlations between talker-specificity and lexical competition effects

Correlational analyses were conducted, as in Experiment 2, using lexical competition data, and d′, β, and RT data from the old/new categorisation task. None of the correlations survived a Bonferroni correction, again suggesting that the time-course of TSEs and lexical competition effects follow independent time-courses.

3.4.3 Discussion

Experiment 3 replicates the pattern of TSEs observed in Experiment 2, showing that these effects are robust and stable across the course of a week after initial exposure to a set of novel nonwords. Experiment 3 also demonstrates that talker-information influences confidence ratings associated with the old/new categorisation task, with more confident responses given to same-talker items at all time-points. Moreover, participants are able to explicitly access information about the study talker, although the accessibility of this information appears to decrease over the week. Interestingly there was also a significant decrease in TSEs in the β data from the old/new categorisation task over the course of a week, as in Experiment 2, although in the current experiment TSEs only approaching significance on Day 8 whereas they were still significant at this time-point in Experiment 2. It is possible that TSEs in the β data decrease over time as participants

begin to realise that talker information does not provide a reliable cue as to whether an item is old or new. On Day 1 talker information may have a large impact on the response criterion level since participants may be at least partially aware of the match between certain voices and words during study. During the test session on Day 1 only half of the items remained in the same voice as study, thus the reliability of talker information as a cue to whether an item is old of new is only 50%. As a result, talker information may be given a smaller weighting in determining the response criterion in later test sessions. In Experiment 3 in particular, where participants were explicitly informed that the test talker would not necessarily be the same as the study talker for each item, talker information no longer affected response biases in the old/new categorisation task on Day 8.

In comparison to TSEs, the pattern of lexical competition effects is less clear. Interactions between day and word-type (test *vs.* control) suggest that lexical competition emerges only on Day 2 after a period of sleep-associated offline consolidation, consistent with Experiment 2. However, when the data were analysed separately for each test session, the lexical competition effects on Day 2 and Day 8 were not significant. One explanation may be that, if as the pattern of data in Figure 3.5b suggests, lexical competition is observed only for same-talker items, then this may limit the power of the analyses since only 12 test basewords were heard in the same voice as the corresponding novel nonwords were studied in. However, it is interesting that significant competition effects were observed in lexical decision using the same number of items. An alternative explanation may be that the null effects in pause detection arose due to the reduced number of filler items in this task compared to the lexical decision task where both word and nonword items were used as fillers. It is possible that participants were more likely to respond strategically or with decision biases in the pause detection task where the link between the novel nonwords and the basewords may have been more apparent.

One final point to note is that the data set, although not showing statistically significant lexical competition effects, do show a pattern of data that is suggestive of talker-specific lexical competition. Given that pause detection is a more online measure of lexical competition than lexical decision, it therefore seems unlikely that the talker-specific lexical competition effects observed in Experiment 2 can be fully accounted for by Luce et al.'s (2003) time-course hypothesis.

3.5 Chapter Summary and Discussion

The experiments described in this chapter show that TSEs and lexical competition effects for novel nonwords follow a different time-course during word learning. TSEs emerge immediately after novel words are studied, and remain constant across the course of a week. Lexical competition effects on the other hand are absent immediately after studying a set of novel nonwords, but emerge one day later following a period of sleep-associated offline consolidation. The different time-courses of these two effects may represent a distinction in storage, with TSEs being more reliant on highly detailed, hippocampally-mediated representations, and lexical competition effects emerging only once neocortical representations have been established.

The fact that TSEs emerge on Day 1, prior to any evidence of integration between existing and novel words suggests that these effects are likely to be primarily mediated by the hippocampal system (according to the CLS framework; McClelland et al., 1995). As noted in Chapter 2, the hippocampus is believed to be important in episodic memory, with sparse coding and pattern-separation providing the ideal conditions for highly-detailed memory representations to be formed. The stability of TSEs over the course of a week in d' measures in the old/new categorisation task in Experiments 2 and 3 suggests that TSEs continue to be driven by the same hippocampal system at all time points. This proposal is consistent with the finding that amnesia following hippocampal damage is often temporally graded (Nadel & Moscovitch, 1997) and can span several years prior to the point in time where damage occurs, suggesting that hippocampal representations continue to be involved in memory for a considerable period of time after initially encountering an item or piece of information.

It could be argued that TSEs were observed at all time points only because talker information was deliberately and strategically encoded during study in order to aid later recognition of the novel nonwords during the old/new categorisation task. However, an additional experiment, not reported in this thesis, addressed this potential confound by using a surprise old/new categorisation task. At the beginning of this experiment participants were told only about the phoneme monitoring and lexical decision tasks. Moreover, participants completed test sessions only on Days 1 and 8, with the aim of minimizing any effects of re-testing and potential confounds associated with using a within-participants design. Despite these changes to the

design of the old/new categorisation task, d′, β, and RT measures still revealed a significant same-talker advantage in the Day 8 re-test, arguing against the suggestion that the stability of the TSEs in Experiments 2 and 3 arose simply dues to strategic encoding of talker-specific information.

However, TSEs may be observed in old/new categorisation at all time-points only because processing of the items in the old/new categorisation task was relatively slow. According to a time-course hypothesis, (Luce et al., 2003; McLennan & Luce, 2005) TSEs are observed only if there is sufficient time to integrate extra-linguistic details into the retrieved phonological representation. RTs in our old/new categorisation task were relatively slow likely due to the fact that the novel and foil nonwords were identical up to the final consonant (*e.g., biscal vs. biscan*), and as such, decisions about whether the item was old or new had to be withheld prior to this point. Notably, RTs in the stem-completion task (Experiment 1), measured from cue onset to voice onset, were also typically quite long, around 1380ms. Thus, a time-course hypothesis appears to be consistent with our findings.

As for the lexical competition effects, the overall pattern of data is consistent with previous studies showing a role for sleep-associated consolidation in the emergence of these effects (*e.g.,* Dumay & Gaskell, 2007). Interestingly, AUC analysis of confidence ratings, collected in the old/new categorisation task, also indicates that participants become more dependent on familiarity mechanisms that are thought to rely on the neocortex and MTL regions surrounding the hippocampus (Brandt et al., 2009; Elfman et al., 2008; Henson et al., 2003; Ranganath et al., 2004), particularly the perirhinal cortex (Aggleton et al., 2005; Davachi et al., 2003), after a period of sleep-associated offline consolidation. This suggestion is consistent with the notion that sleep is important in allowing neocortical representations of newly-learned words to be established and become able to support spoken word recognition. Nonetheless, the pattern of talker-specific lexical competition effects observed in Experiment 2 is somewhat confusing. On the other hand, the fact that talker-specific lexical competition effects were observed at all suggests that activation of neocortical representations must also involve activation of talker-specific details.

Davis and Gaskell (2009) have suggested that activation of hippocampal representation occurs more slowly than activation of neocortical representations.

Thus, talker-specific lexical competition observed when processing is slow may still be consistent with the suggestion that neocortical representations are abstract. In other words the talker-specific lexical competition effects may reflect co-activation of abstract neocortical representations and episodic hippocampal representations. Consistent with this suggestion, previous experiments have only found specificity effects in lexical decision when the task was made more difficult by introducing nonwords that were very word-like (Gonzalez & McLennan, 2007, exp 2; McLennan & Luce, 2005), as was the case with the nonwords used in Experiment 2. The greater the similarity between words and nonwords in the lexical decision task, the more difficult it becomes to make a word/nonword decision, and as a result responses are likely to be slowed.

An alternative explanation may be that both phonological and talker information are stored in long-term memory, but that there are hemispheric asymmetries involved in the processing and storage of these two types of information (e.g., Marsolek, 1999). According to this explanation abstract and episodic representations may co-exist in long-term memory. However, previous studies in both children and adults have shown improvements in memory for phonological information in cued and free recall tasks following a period of sleep-associated offline consolidation (Brown et al., in press; Dumay & Gaskell, 2007; Henderson et al., submitted-b; Tamminen et al., 2010). Thus, if talker information was subject to the same type of consolidation as phonological information, resulting in storage of both types of information in long-term lexical memory, then we might expect similar strengthening of talker information over time. This was not the case. Contrary to this prediction evidence from Experiment 3 actually indicated that talker information became less influential and less accessible in certain tasks as the week progressed. Firstly, TSEs decreased significantly over the course of a week in β data (as reflected by the interaction between day and test-phase talker). Secondly, accuracy in the male/female categorisation task decreased at delayed test points, suggesting that talker information became less explicitly accessible over time.

Together it seems that the complex set of data from Experiments 1 to 3 is most consistent with a model of the lexicon that incorporates aspects of both episodic and abstract representation. Within a CLS framework evidence suggests that hippocampal representations, assumed to be formed rapidly when novel information is encountered, are highly-detailed and episodic in nature. The nature of

representation in the neocortical network is less clear due to the potential co-activation of hippocampal and neocortical networks at delayed test points. There does however seem to be evidence suggesting that talker-information is not stored in long-term memory, although the presence of talker-specific lexical competition effects appears to be somewhat problematic. I will return to this point in Chapter 5.

CHAPTER 4: THE TIME-COURSE OF TALKER-SPECIFICITY EFFECTS FOR EXISTING AND NOVEL WORDS

4.1 Introduction

The experiments reported in Chapter 3 suggest that talker-specific information is retained in memory and can affect recognition of newly-learned words up to one week post-exposure. In contrast Goldinger (1996) found that TSEs declined over the course of a week for existing words in a similar old/new categorisation task (described in the introduction to Chapter 3).[5] There are a number of differences between Goldinger's experiment and our experiments that may account for the different patterns of TSEs observed.

Firstly, the experiments described in Chapter 3, which explored the retention of talker-specific information over a week for novel words used only two talkers of different genders. According to Geiselman and Crawley's (1983) *voice connotation hypothesis*, abstract gender tags may invoke different connotations of words when participants hear speakers of different genders, and thus spoken word recognition may be gender dependent rather than voice dependent (see also Geiselman & Bellezza, 1977). However, as noted in Chapter 1, Palmeri et al. (1993) have demonstrated that participants are sensitive to both within- and between-gender voice changes, indicating that more information is retained than simply gender information. Moreover, Goldinger (1996) found that the time-course of TSEs for existing words was remarkably similar regardless of whether 2, 6, or 10 talkers were heard during study, suggesting that the retention of talker-specific information over the course of a week for novel nonwords is unlikely to be due simply to participants having used gender tags in these experiments to aid memory for the items.

The number of test points completed by each participant also differed between studies with the experiments described in Chapter 3 using a within-participants design in which each participants completed three test sessions; one immediately after study (Day 1), one a day later (Day 2), and one a week later (Day 8). In comparison, Goldinger (1996) used a between-participants design in which each participant completed only one test session, either on Day 1, Day 2, or Day 8 when

[5] Note that Goldinger (1996) did find evidence of TSEs one week later in an identification-in-noise task, as described in the introduction to Chapter 3, although these TSEs were significantly smaller than those observed in the same task when participants were tested immediately after study.

examining TSEs for existing words. It might be predicted that changing the talker for half of the novel nonwords during the immediate test in a within-participants design should have resulted in smaller TSEs for those items on Days 2 and 8 since presumably these items should be represented by two unique traces in memory after the initial test session, each containing different talker information. This was not the case; TSEs were significant at all time-points for novel nonwords despite the use of a within-participants design. Moreover, TSEs for existing words declined over the course of a week in Goldinger's study despite the fact that participants were all trained on Day 1 but were tested only once. Thus, it seems unlikely that the presence or absence of repeated testing, in which half of the items were repeatedly presented in a different voice to study, is able to account for the observed pattern of data.

Alternatively, it may be that using relatively long novel nonwords derived from low frequency basewords resulted in slower processing of the novel items, allowing additional time for talker-specific information to be incorporated into the retrieved representation, as compared to when monosyllabic existing words were heard in Goldinger's study, a suggestion that would be consistent with Luce et al.'s (2003; McLennan & Luce, 2005) *time-course hypothesis*. Nevertheless, if this explanation were correct, TSEs should have been absent at all time points in Goldinger's experiment, not just on Day 8, since the same monosyllabic items were used in each test session.

More likely, the number of exposures to each item during study was an influential factor. Participants in Goldinger's study heard the existing words only once during the experimental prime block before half of the items changed talker in the subsequent target block whereas participants were repeatedly exposed to the novel nonwords 18 times in Experiments 1-3, spoken consistently by a single talker, before completing the old/new categorisation task. In addition, Goldinger exposed participants to 150 existing words whereas only 24 novel nonwords were studied in the experiments reported in Chapter 3. These differences may have served to increase the salience of talker-specific information during the encoding of the novel nonwords, resulting in stronger TSEs immediately after study, and in turn increasing the likelihood that this information would be retained over time.

On the other hand, the different time-courses of TSEs for existing and novel words may have arisen due to the nature of the words themselves and the fact that existing words have pre-established representations in long-term lexical memory

whereas novel words do not. Within an exemplar (or episodic) model of lexical representation there should already be multiple traces of each existing word in memory prior to the study session of an experiment whereas no traces of novel words should exist. Thus, immediately after study, assuming that a unique episodic trace is generated each time a lexical item is encountered, traces acquired prior to the experiment as well as traces established during the study-phase of an experiment should be activated for existing words whereas only traces acquired during the experiment can be activated for novel words. However, most exemplar models assume that the strength with which individual memory traces are activated depends on the similarity of each trace to the input (Gillund & Shiffrin, 1984; Goldinger, 1998; Hintzman, 1986, 1988). As such, traces acquired during the study-phase of the experiment will be activated more strongly during same-talker compared to different-talker test trials, resulting in more accurate old/new categorisation of the studied items heard in the same voice at both study and test.

Nonetheless, it is possible that the same-talker advantage may differ in size for existing and novel words even in the immediate test. In a simulation using MINERVA 2 (an extreme exemplar model in which a unique trace is generated each time an item is encountered regardless of the similarity of that instance/event to previously stored traces) Goldinger (1998) showed that the greater the number of traces stored in the model's lexicon prior to the simulation the smaller the same-talker advantage. In other words, the greater the number of traces activated at test, the smaller the contribution of each individual trace to the retrieved representation, even those traces that were a perfect match to the test probe. The implication of this finding is that an extreme exemplar model such as MINERVA 2 appears to predict that TSEs should be smaller for existing compared to novel words, even when participants are tested immediately after study, due to the greater number and variety of traces in memory for existing words prior to an experiment.

Returning to the different time-courses of TSEs observed for existing and novel words in previous experiments, it is important to consider how an exemplar model might account for these different patterns of data. Exemplar models typically assume that some forgetting occurs over time, either due to trace decay (MINERVA 2; Goldinger, 1998; Hintzman, 1986), or due to retroactive interference (SAM; Gillund & Shiffrin, 1984). It is possible that the decay of traces and loss of talker-specific information over time may account for the decrease in TSEs observed for

existing words in Goldinger's (1996) old/new categorisation task. In support of this suggestion an additional MINERVA 2 simulation (Goldinger, 1998) showed that TSEs for existing words (*i.e.*, words with traces already established in the lexicon prior to the simulation) decreased as the number of forgetting cycles in the simulation increased. Unfortunately this simulation held the number of prior traces stored in memory for each word constant and thus cannot address the question of whether different patterns of TSEs would have been observed for items with more or less traces established in the lexicon prior to simulation of forgetting cycles. Nonetheless, this finding supports the suggestion that trace decay may be important in accounting for the decrease in TSEs over time for existing words (Goldinger, 1996). However, if the decrease in TSEs over time for existing words was simply due to trace decay then a similar decrease in TSEs for novel nonwords should also have been observed in Experiments 2 and 3 since, presumably, talker information would also be forgotten at the same rate for these items. This was not the case in Experiments 2 and 3 where TSEs for newly-learned words were significant at all time-points.

Alternatively trace decay and the number of prior traces stored in memory might interact to produce different patterns of TSEs for existing and novel words. On different-talker trials traces acquired during the study-phase of an experiment will be activated primarily with respect to their phonological match to the test probe since information about the study talker will not match the talker information contained in the test probe. Therefore, even if talker-specific details decay over time the contribution of these study-traces to the retrieved representation of a word on different-talker trials should remain largely unaltered over time. On the other hand, on same-talker trials as traces from the study-phase of an experiment decay and talker-specific details are lost, the contribution of these traces to the retrieved representation of a test item will decrease, allowing other traces that were acquired prior to the experiment to influence the retrieved representation to a greater extent. For novel nonwords that do not have any traces stored in memory prior to the experiment there will be no additional traces to contribute to the retrieved representation, resulting in less 'interference' for these novel items. As such, TSEs

should be maintained for a longer period of time for novel compared to existing words.[6]

Hybrid models may also be able to account for the different patterns of TSEs for existing and novel words observed in previous experiments. Within a hybrid model episodic and abstract representations may co-existing and make different contributions to recognition memory depending on the novelty of an item and/or the delay since an item has been studied. It seems likely that highly-detailed episodic representations are dominant and contribute more to the retrieved representation of recently encountered items when participants are tested immediately after study. Evidence supporting this claim comes from the finding that TSEs were significant immediately after study in both Goldinger's (1996) study as well as in Experiments 1-3. As such, a hybrid model would make the same predictions as an exemplar model with regards to the size of TSEs for existing and novel words in the immediate test-session. Even if a hybrid model assumed that abstract representations contributed to recognition memory immediately after study, abstract representations would only be available for existing words. Lexical competition data from Experiment 2 suggests that more abstract representations of novel nonwords that are capable of engaging in competition within phonologically-similar existing words require a period of sleep-associated offline consolidation in order to become established. Thus, hybrid models would still predict that TSEs should be smaller for existing compared to novel words immediately after study.

At delayed test points recognition in a hybrid model is likely to rely to a greater extent on abstract representations (in combination with episodic representations), with the contribution of each type of representation to recognition memory differing depending on whether the item has a pre-established representation in the abstract subsystem or not. Presumably the contribution of episodic traces will be greater for items that have less well established abstract representations. If this is the case then episodic details should continue to affect recognition of newly learned words for a longer period of time than for existing words.

[6] One interesting (although potentially problematic) point to note here is that Hintzman (1988) chose not to include extra-experimental traces in his MINERVA 2 simulations "on the assumption that they would only have negligible effects on performance on the experimental tasks" (p.528). He argued that contextual information in the retrieval cues essentially reduce the effects of extra-experimental traces to zero.

Nonetheless, it is important to rule out the possibility that the observed differences in the time-course of TSEs for existing and novel words are simply due to the methodological differences described above. The two experiments reported in this chapter were designed to directly compare TSEs for existing and novel words, controlling for the number of items studied, the number of exposures to each item, the number of test points for each participant, and the length of the items. In both experiments participants were exposed to 24 existing words and 24 novel words in a phoneme monitoring task, with each item occurring six times during study. A short maths-based distracter task was then completed, followed by an old/new categorisation task (identical to that used in Experiment 3, including confidence ratings and male/female categorisation judgments in addition to the old/new decision) in which participants heard all 24 existing words and 24 novel words as well as 48 foil items. Experiment 4 examined differences in the size of TSEs for existing and novel items immediately after study whereas Experiment 5 compared the time-course of TSEs for these two types of items, requiring participants to complete the old/new categorisation task at two time points; immediately after study and again after a one week delay. Target and foil items used in the old/new categorisation task were paired such that the items in each pair were morphologically (or pseudo-morphologically) related, differing only in the final syllable. This was true for both existing word pairs (*e.g., coherent-coherence*) and novel nonword pairs (*e.g., anecdent-anecdence*). Matching the existing and novel word-pairs in this manner equated the difficulty of the old/new judgment in terms of phonological similarity between the target and foil items for existing and novel word pairs. Moreover, using morphologically related word pairs minimized the use of semantic information in the old/new categorisation task for existing words since word-pairs such as *coherent* and *coherence* are very similar in their meaning. Likewise, even if participants noticed the link between a novel nonword and its existing baseword (*e.g., anecdent* and *anecdote*) they would be unable to use this information to differentiate between studied and unstudied novel nonwords in the old/new categorisation task. It was hoped that equating the two stimulus sets in this manner would encourage participants to used the same, or at least very similar, strategies when making old/new categorisation decisions to both existing and novel words, allowing a more direct comparison of the time-course of TSEs for the two sets of items.

4.2 Experiment 4

Experiment 4 investigated whether there were differences in the size of TSEs for existing and novel words when participants were tested immediately after study. Based on Goldinger's MINERVA 2 simulation showing that TSEs were smaller for items that had a greater number of prior traces stored in the lexicon it was predicted that TSEs should be smaller for existing compared to novel words in this immediate test. If hybrid models assume that the episodic subsystem is dominant immediately after an item is encountered then the same predictions can be made for these models. Even if hybrid models assume that abstract representations may be activated alongside episodic representations immediately after a set of items has been studied smaller TSEs should still be observed for existing compared to novel words due to the absence of established abstract representations of the novel items.

4.2.1 Method

Participants

Thirty two adults (*age range* = 18 – 41 years, 9 male) recruited from the University of York participated in the experiment. Four additional participants were tested but were removed from the data set prior to analysis; two participants had an error score more than 2.5 *SD* above the mean in the study-phase of the experiment suggesting that the items had not been encoded correctly, and two failed to follow instructions in the male/female categorisation task, reporting the test voice of the items rather than the study voice.

Stimuli

Twenty-four pairs of existing words, consisting of two morphologically related words that differed only in the final syllable (*e.g., coherent-coherence*) were selected for the experiment. Twelve of the word pairs ended in *-ence/-ent* or *-ance/-ant*, and 12 ended in *-ism/-ist* or *-ise/-ize*. All words were bisyllabic (*N* = 13) or trisyllabic (*N* = 35), between 5 and 10 phonemes in length (*M* = 7.75), and had relatively low frequencies (*M* = 3.88, *range* = 1-15) according to the CELEX database (Baayen, et al., 1993).[7]

[7] Goldinger (1996) did not report the frequency of the existing words used in his old/new categorisation task. It is possible that some of the items were high-frequency words that may have been encountered between the study session on Day 1 and final test session on Day 8. If this is the

Twenty-four pairs of novel nonwords were created from the existing words used in Tamminen and Gaskell's (2008) longitudinal study of word learning in adults. The 24 existing words were mono-morphemic, between 6 and 11 phonemes in length (*M* = 8.21), and were low in CELEX frequency (*M* = 3.79, *range* = 2-8). Two novel nonwords were created from each existing item by changing the final vowel and consonant cluster to match the endings of the existing words pairs (*e.g., anecdent-anecdence*) such that 12 novel nonword pairs ended in *-ence/-ent* or *-ance/-ant*, and 12 ended in *-ism/-ist* or *-ise/-ize*. Of the novel nonwords 14 were bisyllabic, and 34 were trisyllabic (Appendix D).

The stimuli were divided into two lists, with one item from each of the 48 word pairs in each list. Thus each list contained 24 existing words and 24 novel nonwords. Word-endings were matched across lists such that each list contained an equal number of words with each morphological-ending. One male and one female talker, both native British English speakers, recorded the items using the recording equipment described in Experiment 1. Stimuli were edited and peak amplitude was normalised using Adobe Audition.

On average, items spoken by the female talker were shorter (existing – *M* = 762ms, *SD* = 96ms; novel – *M* = 790ms, *SD* = 77ms) than items spoken by the male talker (existing – *M* = 809ms, *SD* = 132ms; novel – *M* = 820ms, *SD* = 128ms). A repeated-measures ANOVA, with factors talker (male *vs.* female), list (1 *vs.* 2), and word-type (existing *vs.* novel) indicated that the difference in speech rate between talkers was significant, $F(1,92) = 12.37$, $p < .01$, $\eta_p^2 = .12$. There was however no significant difference in speech rate between lists, $F < 1$, or between existing and novel words, $F < 1$. As such, any differences between the two types of word in the experimental tasks cannot be attributed to differences in time taken for the two talkers to articulate the two sets of items.

Design and Procedure

All tasks in Experiment 4 were completed in a single test session lasting approximately 45 minutes. During the *study-phase* of the experiment each participant was exposed to one list of 48 items (24 existing words and 24 novel

case then it may be one reason why TSEs decreased over the course of a week in Goldinger's study. In order to minimize the possibility that participants would encounter the existing words outside of the experiment (as would also be the case for the novel words) low frequency existing words were selected.

nonwords) in a phoneme monitoring task. Half of the participants were exposed to each list. Within each list 12 of the existing words were spoken consistently by the male talker, and 12 consistently by the female talker. The same was true for the novel nonwords. Study talker was counterbalanced across participants for each individual item. The phoneme-monitoring task was identical to that used in Experiments 1-3 except that each item was heard only six times, once per block.

Following the phoneme monitoring task participants completed the maths-based *distracter task* from Experiment 1 in order to minimize short-term recency effects in later recognition tasks (Goh, 2005). Data from this task are not reported.

In the *test phase* participants heard all 48 existing words and 48 novel nonwords in an old/new categorisation task identical to that used in Experiment 3. For each item participants first judged whether the word was studied (*old*) or unstudied (*new*), then rated their confidence in this decision before being prompted to indicate whether the study voice of the item was male or female. Critically, half of the studied existing words and half of the studied novel nonwords changed talker between study and test. All unstudied items were heard in the same voice as the corresponding studied word from that word-pair. Thus, if *coherent* was spoken by the male talker at test, then the corresponding foil *coherence* was also spoken by the male talker during the test-phase of the experiment.

4.2.2 Results

Study phase

Sixteen participants were exposed to List 1 (5 male) and 16 to List 2 (4 male). The mean error rate in the phoneme monitoring task was 6.3% (*SD* = 3.4%), and the mean RT was 1103ms (*SD* = 243ms). A repeated-measures ANOVA, with factors study-talker (male *vs.* female), word-type (existing *vs.* novel), and list (1 *vs.* 2) showed that for the error data there was a significant main effect of word-type, $F_1(1,30) = 13.21$, $p = .001$, $\eta_p^2 = .31$, $F_2(1,92) = 12.05$, $p = .001$, $\eta_p^2 = .12$, with more errors being made to novel nonwords (*M* = 7.6%; *SD* = 5.2%) than to existing words (*M* = 5.0%; *SD* = 3.7%). However, there was no main effect of study-talker, $F < 1$, and no interaction between word-type and study-talker, $F_1 < 1$, $F_2(1,92) = 1.14$, *ns*. For the RT data there was a main effect of word-type, $F_1(1,30) = 29.38$, $p < .001$, $\eta_p^2 = .50$, $F_2(1,92) = 12.24$, $p = .001$, $\eta_p^2 = .12$, with faster responses to existing words

(*M* = 1072ms; *SD* = 226ms) than novel words (*M* = 1136ms; *SD* = 267ms). However, as with the error data, there was no main effect of study-talker, $F < 1$, nor was there a significant interaction between word-type and study-taker, $F_1(1,30) = 1.33$, *ns*, $F_2 < 1$. Taken together these findings suggest that whilst, unsurprisingly, processing of existing words appeared to be faster and more accurate than that of the novel nonwords, there was no significant effect of study-talker on either error or RT measures in the study task. This is reassuring as it suggests that any differences in memory for the existing and novel words resulting from a change in talker in the old/new categorisation task were unlikely to be due to differences between processing of items spoken in the male and female voices during encoding.

Talker-specificity effects

For each word heard in the old/new categorization task participants made three responses, an old/new judgment, confidence rating, and a male/female decision. Data from each of these responses were analysed separately.

In the *old/new categorisation* task participants responded correctly to 60.3% (*SD* = 7.4%) of the items when making an old/new categorization. Although accuracy was much lower than in Experiments 1-3, participants still performed significantly above chance, $t(31) = 7.72$, $p < .001$. RTs were measured from the onset of the item up to the point at which a button-press response was made. The mean RT was 2321ms (*SD* = 359ms).[8]

For SDT data a repeated-measures ANOVA with factors test-phase talker (same *vs.* different) and word-type (existing *vs.* novel) revealed a significant main effect of test-phase talker for d′ values, $F(1,30) = 9.14$, $p < .01$, $\eta_p^2 = .23$ (Figure 4.1a), indicating that participants were more accurate in categorising same-talker than different-talker items. The main effect of word-type was not significant, $F(1,30) < 1$, nor was the interaction between word-type and test-phase talker, $F(1,30) < 1$, indicating that both existing and novel words showed a same-talker advantage in old/new categorisation.

[8] As in Experiment 3 this response-time is very long and is likely to reflect the high processing demands required to complete this task in which three responses were made to each item.

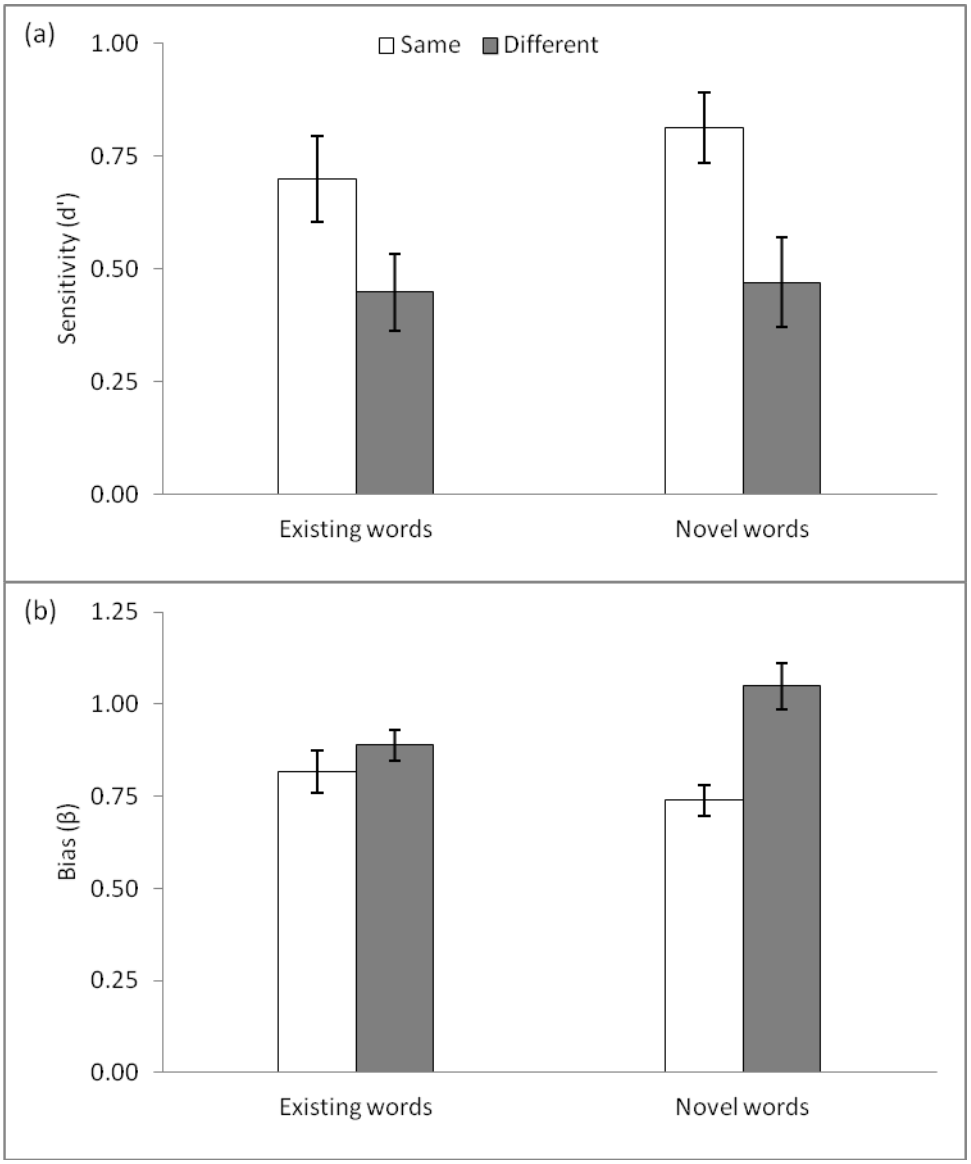

Figure 4.1. (a) Accuracy and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Analysis of β values also revealed a significant main effect of test-phase talker, $F(1,30) = 7.93$, $p < .01$, $\eta_p^2 = .21$, indicating that participants showed significantly different biases when responding to items heard in the same and different voices to study (Figure 4.1b). As in the d′ analysis there was no main effect of word-type, $F < 1$. There was however a significant interaction between test-phase talker and word-type, $F(1,30) = 5.08$, $p < .05$, $\eta_p^2 = .15$, with further analysis revealing that there was a significant main effect of test-phase talker for the novel nonwords, $F(1,30) = 11.81$, $p < .01$, $\eta_p^2 = .28$, but not for existing words, $F < 1$.

As in Experiment 3 *confidence ratings* were used to plot ROC curves, and AUC values were calculated for same- and different-talker existing and novel words separately (Figure 4.2). A repeated-measures ANOVA, with variables test-phase talker (same *vs.* different), and word-type (existing *vs.* novel) revealed a main effect of test-phase talker, $F(1,29) = 10.50$, $p < .01$, $\eta_p^2 = .27$, with greater AUC values for same-talker items indicating a greater contribution of recollection processes to recognition of these items. The main effect of word-type was not significant, $F < 1$, nor was the interaction between test-phase talker and word-type, $F < 1$, supporting the d′ analysis indicating that TSEs were similar in size for existing and novel words.

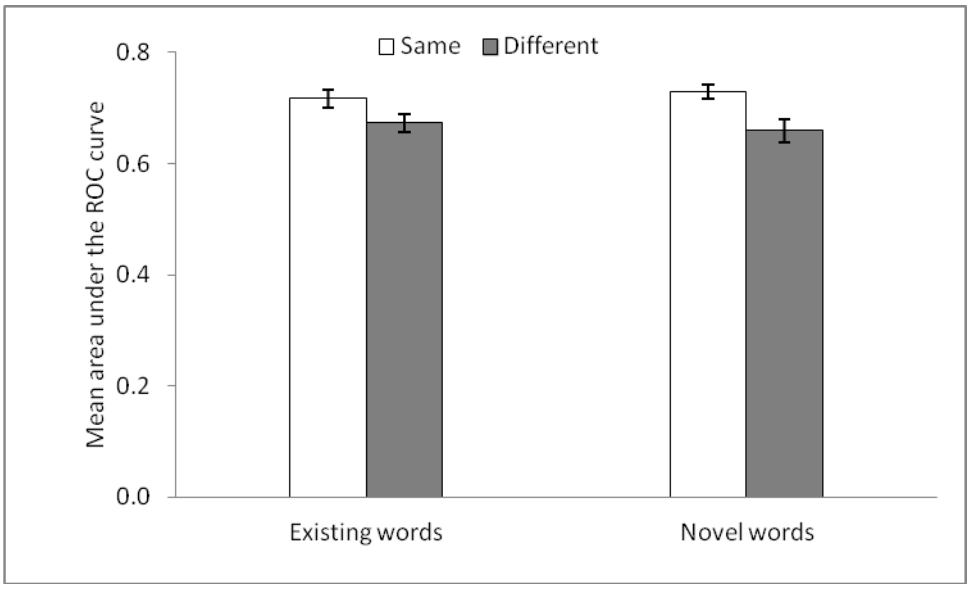

Figure 4.2. Confidence ratings given to old/new categorisation decisions were used to plot ROC curves. Area under the ROC curve was calculated separately for same and different talker existing and novel items. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

In the *male/female categorisation* task only data from studied items were included in the analysis. Overall participants responded correctly 55.1% (*SD* = 5.9%) of the time, significantly above chance, $t(31) = 5.81$, $p < .001$. Although accuracy in this task was relatively low, the percentage correct score is roughly consistent with that observed in a study by Hintzman, Block, and Inskeep (1972) in which participants correctly recalled the study voice 59% of the time, and the case of study

(uppercase *vs.* lowercase) 58% of the time. As in Experiment 3, since participants were instructed to focus on accuracy rather than speed when making a male/female categorisation decision RT latencies were not analysed.

Analysis of the data indicated that there was a main effect of word-type, $F_1(1,30) = 6.99$, $p < .05$, $\eta_p^2 = .19$, $F_2(1,91) = 7.14$, $p < .01$, $\eta_p^2 = .07$, with higher overall performance for existing words (M = 60.8%, SD = 10.4%) compared to novel words (M = 53.8%, SD = 12.8%). There was also, unsurprisingly, a significant main effect of test-phase talker, $F_1(1,30) = 62.00$, $p < .001$, $\eta_p^2 = .67$, $F_2(1,91) = 113.15$, $p < .001$, $\eta_p^2 = .55$, with significantly more correct categorisation decisions being made to items heard in the same voice as study (M = 72.0%, SD 15.7%) than to items heard in a different voice (M = 42.6%, SD = 18.3%). However, the interaction between test-phase talker and word-type was not significant, $F < 1$, indicating that the same-talker advantage was equivalent for existing and novel words (Table 4.1).

Table 4.1. Percentage of correct responses in the male/female categorisation task (studied existing and novel words only), split according to whether the item was spoken in the same or a different talker to study.

	Existing	**Novel**
Same	75.3	68.8
Different	46.4	38.8

4.2.3 Discussion

To summarise, d′ and AUC data revealed a significant same-talker advantage immediately after study. The presence of significant TSEs for novel words immediately after study is consistent with Experiments 1-3, as well as eye-tracking studies by Creel and colleagues (Creel et al., 2008; Creel & Tumlin, 2009, 2011) showing that talker information can affect recognition of recently learned words. Likewise, the presence of TSEs for existing words immediately after study is consistent with a number of previous studies showing a same-talker advantage during recognition of recently studied existing words (Bradlow et al., 1999; Goh, 2005; Goldinger, 1998; Palmeri et al., 1993; Pilotti et al., 2000; Sheffert, 1998). However, the absence of a significant interaction between word-type and test-phase talker in both the d′ and AUC data suggests that TSEs were equivalent in size for existing and novel words at this time point, contrary to predictions made on the basis of Goldinger's (1998) MINERVA 2 simulation showing that the greater the number

of traces stored in the model's lexicon prior to the simulation the smaller the same-talker advantage. Moreover these findings are inconsistent with hybrid models, which may suggest that recognition of existing words should be able to rely on a combination of both episodic and abstract representations immediately after study. Thus, the findings from Experiment 4 appear to contradict the predictions of both exemplar and hybrid models of lexical representation.

One explanation as to why there were no differences in the size of TSEs for existing and novel words in the d′ and AUC data in the current experiment may be that processing of all items was relatively slow, allowing time for talker-specific information to be integrated with information about the phonological form of each words (consistent with Luce et al.'s, 2003, *time-course hypothesis*). A number of factors may have contributed to slow processing of items in Experiment 4. Firstly, requiring participants to make three decisions to each item in the recognition test resulted in slow old/new categorisation decisions. Secondly, using morphologically (or pseudo-morphologically) related word-pairs in the old/new categorisation task meant that an old/new decision could not be made until the final syllable of each word had been heard. Thirdly, all existing words and basewords used to generate the novel nonwords were relatively low in frequency (1-15 according to the CELEX database, Baayen et al., 1993). If Luce and colleagues' (Luce et al., 2003; Luce & Lyons, 1998, 1999; Luce & McLennan, 2005) time-course hypothesis is correct, then slower processing should have allowed more time for episodic details to be integrated into the retrieved representation of all items, resulting in minimal differences between the TSEs observed for existing and novel words.

Interestingly a difference between existing and novel words was observed in the β data, with a significant same-talker advantage emerging only for novel words, consistent with the predictions of both exemplar and hybrid models. This finding suggests that talker information biases the response criterion to a greater extent when making old/new categorization responses to novel words compared to existing words. Nevertheless, it is interesting that smaller TSEs for existing compared to novel words emerged only in the β data, not in the d′ and AUC data. More confusingly, explicit recall of information about the study talker in the male/female categorisation task was in fact better overall for existing words compared to novel words. One possible explanation may be that when a new word is encountered

processing resources must be directed towards encoding and storage of phonological information in order to be able to recognise later instances of the same word. By comparison, if a word already has an established phonological representation in memory then fewer processing resources are likely to be needed to encode and store the phonological form of that word, freeing up processing resources for encoding of additional extra-linguistic details. However, the problem with this explanation is that if talker information was encoded in greater detail for existing words then larger TSEs for existing compared to novel words should also have been observed in our other tasks. As such, better explicit recall of information about the study talker for existing compared to novel words is difficult to account for.

4.3 Experiment 5

Building on Experiment 4, which suggested that TSEs were similar in size for existing and novel words immediately after study in the d′ and AUC data, Experiment 5 examined whether the time-course of TSEs was similar or different for existing and novel words over the course of a week once the two sets of items were equated on variables such as the number of items in each set, the number of exposures to each item during study, and the number of test sessions completed by each participant. Few studies have compared surface-form specificity effects for spoken existing words and nonwords, particularly not with a focus on the time-course of these effects. However, there are a number of studies investigating repetition priming for written words and nonwords that may be informative in this matter. Repetition priming refers to the facilitation seen during word identification as a result of that word having been recently encountered. Studies investigating repetition priming for words and pronounceable nonwords have found significant effects of repetition for both type of word (Feustel et al., 1983; Rueckl, 1990; Salasoo, Shiffrin, & Feustel, 1985; Scarborough, Cortese, & Scarborough, 1977), and also that these repetition effects decreased to roughly the same extent for words and pseudowords between sessions on different days (Salasoo et al., 1985), suggesting that existing and novel words are represented and processed in a similar manner at both immediate and delayed time points. Nevertheless, these studies of repetition priming do not address the time-course of surface-form specificity effects.

As discussed earlier, exemplar models of the lexicon predict that talker-specific information should decrease over time due to trace decay for both existing and novel words. If the rate of trace decay is equivalent for existing and novel words then explicit access to information about the study talker should decrease equally over time for both sets of items. However, differences between existing and novel words may emerge in the old/new categorisation task since existing words are represented by traces acquired both during and prior to the experiment whereas novel words are represented only by traces acquired during the experiment. Thus, TSEs should decrease to a greater extent over the course of a week for existing compared to novel words due to the greater contribution of extra-experimental traces to retrieved representations of existing words once study traces begin to decay. Changes over time are likely to be particularly evident during recognition of existing words heard in the same voice as study since it is these items that are biased to the greatest extent in the immediate test by traces established during study. Loss of talker information as a result of trace decay is likely to have a smaller effect on different-talker test trials since recognition of these items is likely to rely primarily on retrieval of phonological, but not talker-specific, information from recently acquired traces at both immediate and delayed test points.

A similar set of predictions is made by hybrid models of the lexicon. Within a hybrid model it is assumed that whilst episodic representations dominate initially, abstract representations gradually become more dominant over time. Trace decay of episodic representations may be one of the factors driving the change in reliance between the episodic and abstract subsystems. However, whilst existing words should already have robust phonological representations in the abstract subsystem novel words do not. Experiment 2 indicated that abstract representations of novel words that were capable of engaging in lexical competition with phonologically-similar existing words were not established until after a period of sleep-associated offline consolidation. Thus, for novel words at least, recognition could only begin to rely on the abstract subsystem on Day 2. Even after an initial period of sleep-associated offline consolidation the abstract representations of novel words may not be fully established; several periods of offline consolidation may be required to allow a new lexical entry to be established (*e.g.,* Tamminen, 2010). As such, it may be that abstract representations contribute more to recognition of existing compared

to novel words even at delayed test points, and so TSEs should decrease to a greater extent for existing compared to novel words when tested one week after study.

4.3.1 Method

Participants

Forty adults (*age range* = 18-33 years, 10 male) recruited from the University of York participated in the experiment. Eight additional participants were tested but were replaced due to absence from the second test session (2), equipment failure resulting in loss of data (1), failure to follow instructions in the male/female categorisation task, reporting the test voice of the items rather than the study voice (3) or having an error score more than 2.5 *SD* above the mean in the phoneme monitoring task indicating that the items had not been correctly encoded (2).

Stimuli

The stimuli and counterbalancing of stimuli across participants was the same as in Experiment 4.

Design and Procedure

Each participant completed two sessions separated by one week (Day 1 and Day 8). On Day 1 participants completed the phoneme monitoring, maths-based distracter, and old/new categorisation tasks, as in Experiment 4. On Day 8 participants completed only the old/new categorisation task. All tasks and instructions were identical to those used in Experiment 4.

4.3.2 Results

Study phase

Twenty participants were exposed to List 1 (7 male), and 20 to List 2 (3 male). The mean error rate in the phoneme monitoring task was 5.3% (*SD* = 2.1%), indicating that participants were paying close attention to the phonological form of the words during the study phase of the experiment. The mean RT was 1097ms (*SD* = 255ms). A repeated-measures ANOVA, with factors study-talker (male *vs.* female), word-type (existing *vs.* novel), and list (1 *vs.* 2) revealed a significant main effect of word-type in the error data, $F_1(1,38) = 76.78$, $p < .001$, $\eta_p^2 = .67$, with

significantly more errors being made to novel words (*M* = 6.9%, *SD* = 4.1%) than to existing words (*M* = 3.7%, *SD* = 2.7%), although this main effect of word-type was not significant by-items, $F_2(1,92) = 1.42$, *ns.* The main effect of study-talker was non-significant, $F_1(1,38) = 1.11$, *ns*, $F_2(1,92) = 3.43$, $p = .067$, $\eta_p^2 = .04$, as was the interaction between study talker and word-type, $F < 1$. Analysis of RT latencies also revealed a significant main effect of word-type, $F_1(1,38) = 36.19$, $p < .001$, $\eta_p^2 = .49$, $F_2(1,92) = 10.57$, $p < .01$, $\eta_p^2 = .10$, with faster RTs to existing words (*M* = 1071ms, *SD* = 246ms) than to novel words (*M* = 1124ms, *SD* = 267ms). The main effect of study-talker was also significant in the by-participants analysis, $F_1(1,38) = 7.76$, $p < .01$, $\eta_p^2 = .17$ suggesting that responses were quicker to items heard in the male voice at study (*M* = 1089ms, *SD* = 254ms) compared to items heard in the female voice (*M* = 1089ms, *SD* = 263ms). Nevertheless, this main effect was not significant by-items, $F < 1$. In addition, the interaction between study talker and word-type was not significant in either analysis, $F_1(1,38) = 2.53$, *ns*, $F_2 < 1$.

Test-specificity effects

In the *old/new categorisation* task participants responded correctly to 60.0% (*SD* = 4.1%) of the items, a level significantly above chance, $t(38) = 16.80$, $p < .001$. One participant had an error score more than 2.5 *SD* above the grand mean. Data from this participant was removed prior to analysis. The mean RT, measured from word onset up to the point at which a button-press response was made, was 2166ms (*SD* = 365ms).

Analysis of d′ values (Figure 4.3a) using a repeated measures ANOVA with factors test-phase talker (same *vs.* different), day (1 *vs.* 8), and word-type (existing *vs.* novel word) revealed significant main effects of test-phase talker, $F(1,36) = 18.11$, $p < .001$, $\eta_p^2 = .34$, and word-type, $F(1,36) = 1.94$, $p < .01$, $\eta_p^2 = .19$, with higher d′ values for same-talker items and for novel words respectively. The main effect of day was not significant, $F(1,36) = 1.94$, *ns*, suggesting that overall accuracy in the old/new categorisation task did not change across the course of a week. None of the interactions approached significance.

Given the intriguing pattern of data plotted in Figure 4.3a further exploratory analyses were conducted. When the d′ data were analysed separately for existing and novel words the main effect of test-phase talker was significant for both sets of items

(existing – $F(1,36) = 5.43$, $p < .05$, $\eta_p^2 = .13$; novel – $F(1,37) = 14.60$, $p < .001$, $\eta_p^2 = .28$). Likewise, both sets of items showed non-significant main effects of day (existing, $F(1,36) = 2.04$, *ns;* novel words, $F < 1$) as well as non-significant interactions between day and test-phase talker (existing – $F < 1$; novel – $F < 1$). These findings suggest that overall performance in the old/new categorisation task remained stable over the course of a week for both existing and novel words and that TSEs did not change in size for either type of item between Days 1 and 8 respectively. However, when the data from each word-type were analysed separately on each day there were significant main effects of test-phase talker for both existing and novel word on Day 1, and for novel words on Day 8 (Day 1, existing – $F(1,37) = 6.09$, $p < .05$, $\eta_p^2 = .14$; Day 1, novel – $F(1,37) = 8.81$, $p < .01$, $\eta_p^2 = .19$; Day 8, novel – $F(1,37) = 6.09$, $p < .05$, $\eta_p^2 = .14$). In comparison, the main effect of test-phase talker was non-significant for existing words on Day 8, $F(1,36) = 2.23$, *ns*, suggesting that there may be a subtle change in the pattern of d' data for existing words on Day 8.

There are two further pieces of evidence suggesting that existing words may be processed somewhat differently on Day 8 compared to Day 1. Firstly, when same- and different-talker items from each word-type were analysed separately the main effect of day was marginally significant only for same-talker existing words, $F(1,37) = 3.49$, $p = .07$, $\eta_p^2 = .09$, suggesting that there was a decline in d' scores for same-talker existing words over the course of a week. Secondly, when the data were analysed separately for same- and different-talker items individually on each day there was a significant main effect of word-type only for same-talker items on Day 8, $F(1,37) = 6.76$, $p = .013$, $\eta_p^2 = .16$, with lower d' scores for same-talker existing words compared to same-talker novel words on Day 8. There was no main effect of word-type for different-talker items Day 8, $F < 1$. Together these findings point towards a marginal decline in d' scores for same-talker existing words over the course of a week, with this decline resulting in a significant difference in d' scores between same-talker existing and novel words on Day 8 as well as non-significant TSEs for the existing words themselves at this delayed test-point. These findings are intriguing and potentially point to a subtle difference in processing of existing and novel words on Day 8. However, given that all of the critical interactions were non-significant in the main analysis these findings must be interpreted with caution.

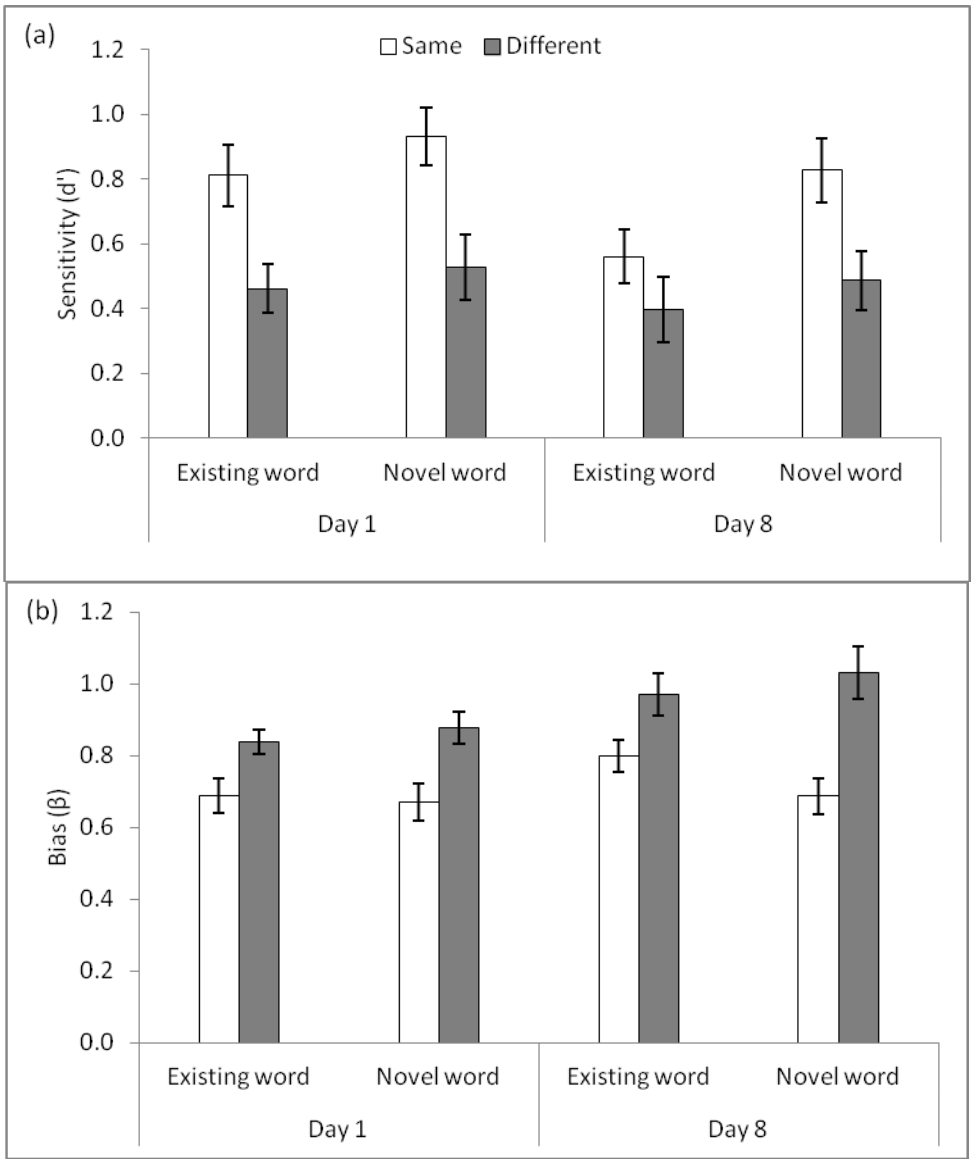

Figure 4.3. (a) Accuracy and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Analysis of the β values (Figure 4.3b) revealed significant main effects of test-phase talker, $F(1,36) = 26.50$, $p < .001$, $\eta_p^2 = .42$, and day, $F(1,36) = 4.89$, $p < .05$, $\eta_p^2 = .12$, but a non-significant main effect of word-type, $F < 1$. There were significant two-way interactions between test-phase talker and day, $F(1,36) = 3.33$, $p = .076$, $\eta_p^2 = .09$, and between test-phase talker and word-type, $F(1,36) = 4.08$, $p = .051$, $\eta_p^2 = .10$. In order to explore these interactions further same- and different-talker items were analysed separately. The main effect of day was significant for

different-talker, $F(1,36) = 7.33$, $p = .01$, $\eta_p^2 = .17$, but not same-talker items, $F < 1$. In comparison the main effect of word-type began to approach significance for same-talker, $F(1,37) = 2.93$, $p = .095$, $\eta_p^2 = .07$, but not different-talker items, $F(1,36) = 1.05$, *ns*.

Note, when the data from each test-session were analysed separately (in order to explore the interaction between day and test-phase talker) the main effect of test-phase talker was significant at both time points (Day 1 – $F(1,37) = 11.95$, $p = .001$, $\eta_p^2 = .24$; Day 8 – $F(1,36) = 21.91$, $p < .001$, $\eta_p^2 = .38$. Likewise, when the data from each word-type were analysed separately (in order to explore the interaction between word-type and test-phase talker) the main effect of test-phase talker was significant for both existing, $F(1,36) = 7.93$, $p < .01$, $\eta_p^2 = .18$, and novel words, $F(1,37) = 25.57$, $p < .001$, $\eta_p^2 = .41$.

Analysis of *confidence ratings* using AUC values derived from ROC curves revealed main effects of day, $F(1,38) = 10.94$, $p < .01$, $\eta_p^2 = .22$, test-phase talker, $F(1,38) = 21.79$, $p < .001$, $\eta_p^2 = .36$, and word-type, $F(1,38) = 4.41$, $p < .05$, $\eta_p^2 = .10$, with higher AUC values on Day 1, for same-talker items, and for novel words respectively. The only interaction that approached significance was the three-way interaction between day, test-phase talker, and word-type, $F(1,38) = 3.33$, $p = .076$, $\eta_p^2 = .08$. In order to explore this interaction the data from each word-type were analysed separately. The main effect of test-phase talker was significant for both sets of items (existing – $F(1,38) = 13.04$, $p = .001$, $\eta_p^2 = .26$; novel – $F(1,38) = 11.19$, $p < .01$, $\eta_p^2 = .23$). In comparison the main effect of day was significant only for existing words, $F(1,38) = 11.84$, $p = .001$, $\eta_p^2 = .24$, not for novel words, $F(1,38) = 2.66$, *ns*, suggesting that there was a change in reliance on recollection and familiarity processes over the course of a week only for existing words. Nonetheless, the interaction between test-phase talker and day was non-significant for both sets of items (existing – $F < 1$; novel – $F(1,38) = 2.53$, *ns*) indicating that the size of the same-talker advantage remained stable over time for both existing and novel words.

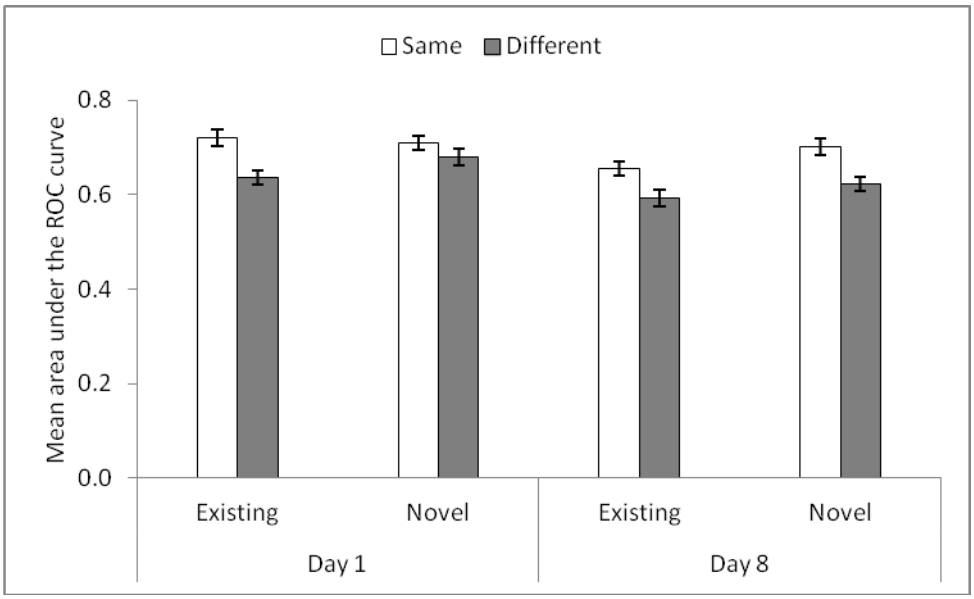

Figure 4.4. Confidence ratings given to old/new categorisation decisions were used to plot ROC curves. Area under the ROC curve was calculated separately for same and different talker existing and novel items in each test session. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

As in previous experiments only data from studied items were included in the analysis of *male/female categorization* data. For studied words, participants responded correctly to 59.0% (*SD* = 6.8%) of the items, a level of performance significantly above chance, $t(39) = 8.19$, $p < .001$. The mean RT, measured from the onset of the words '*male – female*' on screen until a button-press response was made, was 764ms (*SD* = 450ms). Since participants were instructed to focus on accuracy, not speed of response, when making the male-female categorization decision the RT data was not analyzed further.

A repeated-measures ANOVA with day (1 *vs.* 8), test-phase talker (same *vs.* different), and word-type (existing *vs.* novel) revealed that there was a main effect of day, $F_1(1,38) = 8.62$, $p < .01$, $\eta_p^2 = .19$, $F_2(1,92) = 16.01$, $p < .001$, $\eta_p^2 = .15$, with participants responding significantly more accurately on Day 1 (*M* = 62.0%, *SD* = 10.4%) than on Day 8 (*M* = 56.1%, *SD* = 9.8%). There was also a significant main effect of test-phase talker, $F_1(1,38) = 71.20$, $p < .001$, $\eta_p^2 = .65$, $F_2(1,92) = 384.72$, $p < .001$, $\eta_p^2 = .81$, indicating that information about the study talker was recalled more accurately for same-talker compared to different talker items. There was however no main effect of word-type, $F < 1$, indicating that there was no difference

overall between performance in the male/female categorisation task for existing and novel words. The only significant interaction was between word-type and test-phase talker, $F_1(1,38) = 10.89$, $p < .01$, $\eta_p^2 = .22$, $F_2(1,92) = 10.50$, $p < .01$, $\eta_p^2 = .10$. In order to explore this interaction further same- and different-talker items were analysed separately. There was a significant main effect of word-type for same-talker items, $F_1(1,38) = 10.39$, $p < .01$, $\eta_p^2 = .22$, $F_2(1,92) = 4.95$, $p < .05$, $\eta_p^2 = .05$, with greater accuracy for same-talker novel nonwords than existing words. This main effect of word-type was also marginally significant for different-talker items, $F_1(1,38) = 3.08$, $p = .087$, $\eta_p^2 = .08$, $F_2(1,92) = 4.52$, $p < .05$, $\eta_p^2 = .05$. However, inspection of the data in Table 4.2 reveals the opposite pattern of data, with accuracy being slightly higher for existing compared to novel words for different-talker items.

Table 4.2. Percentage of correct responses in the male/female categorisation task in each test session (studied existing and novel words only), split according to whether the item was spoken in the same or a different talker to study.

	Existing		**Novel**	
	Day 1	Day 8	Day 1	Day 8
Same	75.6	71.4	80.4	76.3
Different	49.8	39.4	42.1	37.7

4.3.3 Discussion

As in Experiment 4 the d′ and AUC data from Experiment 5 revealed significant same-talker advantages that did not interact with word-type, suggesting once again that TSEs were similar in size for existing and novel words. However, in contrast to Experiment 4 participants were significantly more accurate overall in making old/new judgments for novel compared to existing words. Likewise, AUC scores were higher overall for novel nonwords relative to existing words. It is possible that these effects arose due to the semantic as well as phonological overlap between target and foil existing words compared to the phonology-only overlap between target and foil novel words.[9] If this is the case then old/new categorisation decisions are likely to have been easier for novel compared to existing words, resulting in higher d′ and AUC scores for these items. Nonetheless, the absence of

[9] It is however possible that semantic information associated with the basewords from which the novel nonwords were derived may have affected old/new categorisation decisions for novel nonwords also.

these effects in Experiment 4 suggests that this cannot be the full explanation otherwise higher d′ and AUC scores should also have been observed for novel words in the previous experiment. An alternative explanation may be that the higher overall scores for novel words in the d′ and AUC data stem from the combination of scores across immediate and delayed test sessions in the Experiment 5 analyses. In other words, when only scores from the immediate test are included in the analysis (Experiment 4) there are no differences between existing and novel words. In comparison, when scores from the immediate and delayed test sessions are combined (Experiment 5) there is an overall advantage for novel words. This account would suggest the overall advantage for novel words in Experiment 5 stems from higher d′ and AUC scores for novel compared to existing words in the Day 8 test session, and would add additional support to the suggestion that there may be subtle differences between existing and novel words when participants are tested a week after initially studying the items.

With regards to the time-course of TSEs, the d′ data revealed a non-significant main effect of day suggesting that overall performance in the old/new categorisation task did not decrease over the course of a week. In comparison, overall AUC scores did decrease across the course of a week, suggesting that there was a gradual change in reliance between recollection and familiarity mechanisms. Interestingly, further analyses indicated that the change in reliance between recollection and familiarity mechanisms over the course of a week was significant only for existing words. This latter finding is somewhat consistent with findings from the further exploratory analyses conducted on the d′ data which revealed subtle differences between existing and novel words. Specifically, old/new categorisation of same-talker existing words became marginally more error-prone in the delayed compared to the immediate test session, resulting in a significant difference in d′ scores for same-talker existing and novel words in the delayed test, and non-significant TSEs for existing words at this delayed test point. However, whilst these effects in the d′ data are in the predicted direction, and are consistent with the suggestion that existing words also showed a decrease in the contribution of recollection processes to recognition memory over time, it is difficult to determine their importance given the lack of significant interactions in the main d′ analysis. In particular, the lack of a significant interaction between day and test-phase talker in both the d′ and AUC analyses indicates that

TSEs remained stable between the immediate and delayed test sessions for both sets of items.

The maintenance of significant TSEs for novel nonwords over the course of a week in the d′ and AUC data is consistent with Experiments 2 and 3, reported in Chapter 3. In contrast, the stability of TSEs for existing words over the course of a week is at odds with Goldinger's (1996) study showing that TSEs decreased significantly in size in a similar old/new categorisation task. Even Goldinger's identification-in-noise task (reported in the same paper) showed a significant decrease in TSEs over the course of a week even though these TSEs were still statistically significant in the Day 8 test session. As suggested in the discussion of data from Experiment 4 one explanation for the similar-sized TSEs for existing and novel words in the immediate test may be that processing is relatively slow in the old/new categorisation task, providing additional processing time for episodic details to be integrated into the retrieved representation of all items at test. It is possible that slow response-times in the Day 8 session may have had a similar effect, and that this may be why TSEs appear to be roughly equivalent in size for both sets of items even after a week-long delay. Alternatively, TSEs may be more stable for existing words in Experiment 5 compared to Goldinger's (1996) study due to the smaller number of items studied, and/or the repetition of each item during the study-phase of the experiment. These manipulations may have served to increase the salience of talker-specific information in Experiment 5. Moreover, the use of relatively low frequency words is likely to have minimized the chance of participants hearing the studied items in the delay between immediate and delayed test sessions.

Despite the apparent stability of TSEs in the d′ and AUC data across the course of a week, data from the male/female categorisation task revealed poorer recollection of information about the study talker of each item at the delayed test point. Importantly, day did not interact with either test-phase talker or word-type indicating that the decrease in performance over time was equivalent for both existing and novel words regardless of whether the item was heard in the same or a different voice to study. Thus, the male/female categorisation task indicates that episodic traces of existing and novel words decay over time at a similar rate.[10]

[10] Note, male/female categorisation data in Experiment 4 indicated that explicit recall of information about the study talker was significantly better overall for existing compared to novel words. This was not the case in Experiment 5.

4.4 Chapter Summary and Discussion

To summarise, there appears to be evidence of robust TSEs for both existing and novel words immediately after study, consistent with numerous previous studies examining TSEs for novel (Creel et al., 2008; Creel & Tumlin, 2009, 2011) and existing words (Bradlow et al., 1999; Goh, 2005; Goldinger, 1998; Palmeri et al., 1993; Pilotti et al., 2000; Sheffert, 1998). One week later TSEs were maintained for novel words, consistent with Experiments 2 and 3, and there also appeared to be evidence that TSEs remained stable over time for existing words. This latter finding is inconsistent with previous work by Goldinger (1996), as discussed above, but is somewhat consistent with studies indicating that repetition priming effects decrease to roughly the same extent for existing and novel words over time (*e.g.*, Salasoo et al., 1985). Moreover, the fact that source memory for the study talker of each item appeared to decay at the same rate for existing and novel words further supports the finding that the time-course of TSEs did not differ between existing and novel words.

Nevertheless, data from the confidence ratings task in Experiment 5 revealed a significant change in reliance between recollection and familiarity mechanisms only for existing words. Likewise, further analysis of the d′ data revealed interesting differences between existing and novel words that pointed towards subtle differences in the processing of same-talker items at the delayed test-point. This tentative pattern of findings would be consistent with the suggestion of both episodic and hybrid models, which both predict that TSEs should be smaller for existing compared to novel words at the delayed test point. Exemplar models assume that this difference should arise due to the presence of extra-experimental traces for existing but not novel words. Hybrid models predict that a difference between existing and novel words should arise due to the presence of robust, pre-established abstract representations of the existing but not the novel words. However, the lack of significant interactions in the main d′ analysis is problematic. Thus, the data cannot be taken as strong support for either exemplar or hybrid models.

As suggested above the absence of a significant difference in the size of TSEs for existing and novel words in both test sessions may be due to the relatively slow processing observed in the old/new categorisation task. This slow processing may have arisen due to participants being required to make three responses to each word

heard, to the use of relatively low frequency items, and/or to the use of highly-similar, morphologically related distracters in the old/new categorisation task. Thus it remains possible that if high rather than low frequency existing words had been studied and the distracter items in the old/new categorisation task had not been morphologically related to the studied words/nonwords a difference between the size of TSEs for existing and novel words may have been observed.

In light of this potential problem with the design of our stimuli the significant change over time in AUC scores for existing but not novel words in Experiment 5 is of particular importance. The finding that AUC scores decline significantly over time for existing words suggests that the contribution of familiarity processes to recognition memory increases and the contribution of recollection processes decreases at the delayed test point for existing words. If we assume that recollection is dependent on activation highly-detailed episodic representations within the hippocampus whereas familiarity processes may depend more on abstract representations that rely on the neocortex and MTL regions surrounding the HC (Brandt et al., 2009; Elfman et al., 2008; Henson et al., 2003; Ranganath et al., 2004) then the pattern of AUC data appears to be most consistent with a hybrid model of lexical representation.

It is not necessary to assume that episodic representations are lost as a result of increasing reliance on abstract codes. In fact, data from the male/female categorisation task in Experiment 5 suggest that episodic representations must be maintained for at least one week after initial exposure to an item in order to support above-chance recall of information about the study talker of each item. Evidence supporting the suggestion that episodic representations are maintained for a considerable period of time after initial exposure to an item or event comes from both human and animal studies showing temporally-graded amnesia following hippocampal lesions (see Nadel & Moscovitch, 1997, for a review). Moreover, Walker and Stickgold (2010) have recently suggested that effective integration of information from episodic memory traces into long-term memory may require several nights of sleep in order to be optimal. If this is the case, then episodic memory representations must be maintained for an extended period of time after the item is initially encountered in order to allow successful integration of new and old information, although Experiment 5 provides evidence that these traces also appear to decay gradually over time.

The CLS approach (McClelland et al., 1995) offers one framework in which episodic and abstract representations may co-exist. As described in Chapter 2 the hippocampal system within this framework provides an ideal brain region to support the binding together of information from different sensory inputs to form highly-detailed episodic representations since a large number of cortical areas, as well as almost all association areas, have projections that converge within the hippocampus (Abutalebi et al., 2007; Mayes & Montaldi, 2001; Munoz & Insausti, 2005; Suzuki & Eichenbaum, 2000). If recognition of both studied existing and studied novel items relies primarily on hippocampally mediated representations immediately after study then detailed episodic information about each item should be available and should influence recognition of recently studied items, resulting in significant TSEs for both types of item. In order to make an old/new judgment for existing items at the delayed test point episodic representations must again be accessed since both the studied and unstudied existing words will both have been encountered prior to the experiment. Thus, an old/new judgment cannot be made on the basis of whether participants have a stored abstract phonological representation of the word in long-term memory or not, as may be the case for novel words (although note that all foil nonwords were also encountered during the Day 1 test session, and so may also have become weakly established within the abstract subsystem). Therefore, accurate old/new categorisation of existing words must remain dependent on the hippocampally-mediated episodic representations even at the delayed test point. This may be one reason why significant differences in the size of TSEs for existing and novel words were not observed in the main d′ analysis.

Evidence supporting the suggestion that the hippocampal subsystem must remain involved in making old/new recognition judgments at all time points comes from a study by Holdstock et al. (2002) who compared forced-choice and old/new recognition performance in amnesic patient YR who had a selective hippocampal lesion. YR was able to differentiate between visually-similar studied and unstudied pictures in a two-alternative forced choice, but not in an old/new recognition task where each item was presented separately. The authors argued that following a hippocampal lesion the medial temporal cortex should still be able to generate a familiarity signal associated with each test picture. When two similar pictures are presented together, as in a two-alternative forced-choice task, it is possible

(assuming that studied pictures generate stronger familiarity signals) to determine which of the two item generates a stronger familiarity trace, and thus to differentiate between studied and unstudied items. However, when the two visually-similar pictures are presented separately, as in the old/new categorisation task, both items generate a familiarity signal that is relatively strong, and so both items are likely to be classified as old. In order to differentiate between two highly similar items that are presented separately (such as the word-pairs used in Experiments 4 and 5), it is necessary that the hippocampal systems, assumed to be responsible for recollection (Davachi, Mitchell, & Wagner, 2003; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Ranganath et al., 2004), is involved in the recognition process. Recollection processes may be particularly important in differentiating between studied and unstudied items in Experiments 4 and 5 since evidence suggests that morphologically-complex items are likely to be represented in a decomposed format in the abstract subsystem (Rastle & Davis, 2008).

There are now a large number of studies showing that visually-presented, morphologically-complex words are decomposed into their morphological components at an early stage of processing, and that this is the case regardless of whether or not the whole word-form bears any semantic relation to the stem word. For example masked priming studies have shown that semantically transparent words such as *teacher* prime their root *teach* to the same extent that semantically opaque words such as *corner* prime their root *corn* even though *corner* and *corn* are not semantically related (see Rastle & Davis, 2008, for a review). This decomposition process appears to be robust to orthographic alterations such as ‘e’ deletion (*adorable-adore*), or doubling of the consonant (*drummer-drum*) at the morpheme boundary (McCormick, Rastle, & Davis, 2008). Words sharing a bound morpheme (*deflate-inflate*) also prime each other (Forster & Azuma, 2000), suggesting that all visually-presented morphologically-complex words are represented in a decomposed form within the lexicon regardless of whether the item can be decomposed into whole morphemes or morphemes that cannot stand on their own as words. Evidence suggests that auditorily-presented morphologically-complex words are also decomposed at an early stage of processing (Sedin, 2006), indicating that the morphologically complex existing words encountered in Experiments 4 and 5 are likely to have been represented in a decomposed format. However, given evidence suggesting that the hippocampus is involved in binding information into

coherent representations that are sensitive to even small changes in the input (*e.g.*, Bakker et al., 2008), it seems unlikely that the morphologically-complex studied words will be represented in a decomposed form within the hippocampal memory system. Rather, it seems more likely that items in the neocortical system will be represented in a morphologically decomposed format given the distributed, overlapping nature of representation within this subsystem (McClelland et al., 1995).

Interestingly, Longtin & Meunier (2005) have demonstrated that morphologically complex pseudowords such as *rapidifier* primed the root word *rapide* just as much as morphologically complex existing words such as *rapidement*, suggesting that it is the morphological structure of the word that drives the decomposition process, not the semantic legality of the stem-affix combination. However, Longtin & Meunier found that when an existing stem was combined with an ending that was not an affix (*e.g.*, *-uit* in French), the stem *rapide* was not primed by the combination of stem and non-affix, *rapiduit*, suggesting that morphological decomposition occurs only when both the stem and affix are legal morphemes within the language. If this is the case, then our novel words, which were composed of a novel stem derived from an existing but monomorphemic word plus an existing affix may not have been decomposed into stem and affix. However, Lindsay, Sedin, and Gaskell (2012) have demonstrated that participants are able to decompose novel items into novel stems plus past tense inflections. Participants in this study were exposed to spoken novel words such as *confal*. At test the novel words were heard in a phoneme categorisation task in which the final phoneme of the word ranged along a 9-token continuum from /t/ to /d/ (*e.g.*, *confald*). It was predicted that if participants had stored the novel stem and integrated this novel phonological representation with existing lexical knowledge, more /d/ than /t/ responses should be made in the phoneme categorisation task, reflecting the fact that *confal + /d/* forms the past tense of *confal*, and thus is a more appropriate lexical interpretation of the input than *confal + /t/*, which produces a new word entirely (Ganong, 1980). This is exactly what was found, suggesting that the novel phonological forms were stored immediately after exposure, and were integrated rapidly with existing knowledge about past tense inflection. As such it seems plausible that existing knowledge of the affixes used in the present study may have allowed participants to decompose the morphologically-complex novel nonwords into stem and affix and to store the novel items in this manner in the neocortical system. However, given the differences

between Longtin and Meunier's findings and those of Lindsay et al. it remains unclear whether the novel items were represented as composite wholes or in a morphologically decomposed format in the neocortical system of a hybrid CLS framework. Whilst the use of novel stems in the current experiment raises many questions about how novel morphemes are acquired (see Merkx, Rastle, & Davis, 2011, for experiments examining the acquisition of novel affixes), and at what point these novel stems are treated like existing morphemes within the lexicon, further experiments exploring these questions are, unfortunately, beyond the scope of this thesis.

To conclude, a hybrid model of lexical representation appears to offer the most parsimonious account of the data from Experiments 4 and 5. Due to the nature of the stimulus set and design of the old/new categorisation task it must be assumed that old/new categorisation decisions must rely on the episodic subsystem within a hybrid model at all time points, irrespective of whether the item is existing and novel. In contrast the confidence ratings task appears to suggest that there is an increase in the contribution of the familiarity processes to recognition of existing words at delayed test points, suggesting that abstract representations may also be activated alongside episodic representations at this delayed test point.

CHAPTER 5: THE ROLE OF TALKER VARIABILITY DURING WORD LEARNING

5.1 Introduction

Experiments 1-5 indicate that adults retain talker-specific information in memory for novel words both immediately after study, as well as up to one week later. However, it may be that although highly detailed representations of words are formed initially, as an item is heard in a greater variety of contexts, and spoken by a wider range of people, its representations may become more abstract, either at retrieval (as predicted by exemplar models) or in lexical memory (as predicted by hybrid models). In order to explore this possibility the two experiments reported in this chapter examined whether talker variability during study affected the time-course of TSEs for newly-learned words.

There are a number of studies suggesting that variability in the input during training may have beneficial effects on learning. Developmental studies have demonstrated that when 7.5 month olds are exposed to words spoken by only one talker they do not show recognition of that word when later spoken by a talker of the opposite gender (Houston & Jusczyk, 2000). Similarly, infants do not generalise across instances of the same words spoken in different affects (Singh, Morgan, & White, 2004). However, when infants of the same age were exposed to words spoken in multiple affects during training they later recognised these words when exposed to them in a novel affect, and were also able to differentiate the target word (*e.g. bike*) from a phonologically similar distracter (*e.g. dike*) (Singh, 2008). These findings suggest that exposing infants to multiple different instances of a word can result in a more robust and more abstract representation of that word.

Further evidence supporting this suggestion comes from research investigating learning of novel word–novel object pairs in 14 month olds. Infants exposed to novel words spoken by a single talker did not differentiate between trials in which the habituated novel object was presented with the trained word (*e.g. buk*) or with a minimal-pair distracter (*e.g. puk*). However, infants who were exposed to the novel words spoken by multiple talkers during study were able to differentiate between the minimal-pair novel words (Rost & McMurray, 2009). The authors suggested that infants were better able to learn minimal pairs when the exposure phase contained multiple talkers because the increased variability in the input allowed infants to

'home-in' on the more stable and invariant aspects of the input more effectively. That is, as children were exposed to more varied productions of a word they began to recognise which details in the input were most important, and adjusted their representations accordingly, allowing them to recognise the same word across different speakers (Newman, 2008).

Even in older children talker variability can have beneficial effects during word learning. Richtsmeier, Gerken, Goffman, & Hogan (2009) familiarised 3 to 4 year old children with novel nonwords paired with pictures of 'funny animals'. Some nonwords were heard only once during the passive exposure phase of the experiment whilst other items were heard 10 times (the authors termed this variable *experimental frequency*). When all items were spoken by a single talker during exposure there were no effects of experimental frequency. However, when items were spoken by ten different talkers during exposure, the nonwords that had been heard ten times were subsequently produced quicker than those items that had only been heard once during the exposure phase. Again, this finding suggests that increased variation in the input may facilitate the formation of a more stable representation, which aids later production of that word.

Interestingly, recent research indicates that perhaps not all forms of variability in the input aid word learning. Using the same switch task as described above (in Rost & McMurray, 2009), Rost and McMurray (2010) found that 14 month old infants were unable to discriminate between minimal pair novel words (*e.g. buk* and *puk*) when exposed to those items spoken by a single talker, but with variable voice onset times (VOT) across tokens. Nor were they able to differentiate between the minimal pairs when exposed to the items spoken by a single talker but with multiple sources of variability in the voicing cues (VOT, F0 transition, and burst amplitude). The infants were however able to discriminate between the minimal pairs when exposed to the items spoken by multiple talkers but with a fixed VOT across talkers. These findings are interesting as they suggest that talker-specific information, but not information about fine phonetic details, can be used to aid the development of stable lexical representations in infants.

One limitation of all of the developmental studies described above is that testing always occurred immediately after exposure to the novel items. Thus, it is not clear whether there would be beneficial effects of talker variability if children were tested at a later point in time. Jusczyk, Pisoni, and Mullennix (1992) addressed this

question in two-month old infants using a high amplitude sucking technique. Infants were habituated to a single word (*dug* or *bug*) spoken by either a single talker, or by a set of six talkers (three male and three female). After a two minute delay, during which infants were shown a series of colourful pictures but were not exposed to any further auditory stimuli, infants detected a phonetic change between the two words only when a single talker had been heard during the habituation phase of the experiment, but failed to detect this change when multiple talkers had been heard during habituation. Most strikingly, infants also failed to detect a change between *dug* and *bug* after being exposed to multiple tokens of one of these words all spoken by a single talker during habituation. The authors argue that any type of variability in the input affects the way that two-month old infants remember words that were previously presented during habituation. Taken together, these developmental studies suggest that whilst variability in the input during exposure to a novel word may benefit recognition when tested immediately after training, variability may have detrimental effects on recognition when the test occurs after a short delay. Nevertheless, given that the effect of variability on retention of a novel word was only tested in the youngest group of infants (two-month olds) it is unclear whether the apparent detrimental effect of variability would be found in older age groups who have greater cognitive capacity and executive control, and thus may be better able to cope with variability in the speech input and to use this information to their advantage by homing in on stable structures within the input.

Interestingly, in adults there is evidence that exposing native Japanese speakers to the English /r/-/l/ contrast (a contrast that does not existing in Japanese) spoken by multiple talkers enabled participants to learn the distinction between these two phonemes (Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991), with participants showing improvements in both perception and production of this novel phonetic contrast up to three months post-training (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994). It is important to note that Lively et al. (1993) did not find improvements in discrimination of the English /r/-/l/ contrast when only a single talker was used during training. More importantly, when only a single training talker was heard participants did not generalise their ability to discriminate between /r/ and /l/ when presented with these phonemes in the context of unstudied words. Participants exposed to five talkers during training did show this generalisation, although

notably, generalisation was better when the generalisation words were spoken by one of the training talkers compared to when they were spoken by a novel, unfamiliar talker (Lively, et al., 1994; Logan, et al., 1991), suggesting that talker-specific information was encoded and stored during the training phase of these experiments despite variability in the input. On a similar note, Bradlow and Bent (2008) have demonstrated that native English listeners are able to achieve talker-independent adaptation to Chinese-accented English if exposed to multiple talkers of Chinese-accented English during training, but not if exposed to only a single talker. Sadakata and McQueen (2011) also note the benefits of high-variability training for Dutch speakers learning a Japanese geminate-singleton fricative contrast, and Clopper and Pisoni (2004) demonstrate that exposure to more variable input aids the formation of more robust perceptual categories associated with different regional dialects. Together these findings suggest that variable input is vital in order for adults to form robust, abstract perceptual categories.

Although the studies described above are somewhat different in their questions and methodologies to the word learning studies described in Experiments 1-5, variability in the input during learning appears to result in more robust representations of both novel words (in developmental studies), and novel phoneme categories (in adult studies). These findings are supported by research showing that variability also aids learning of grammatical features of artificial languages (Gomez, 2002). Likewise, in the visual domain there is evidence suggesting that the greater the number of pictures used to create an 'averaged' face of a famous person, the faster and more accurately that image is identified (Burton, Jenkins, Hancock, & White, 2005).

One way to explain the beneficial effects of variability during learning may be to assume, consistent with exemplar models of the lexicon, that as an increasing number of episodic traces of a novel word are stored in memory, the retrieved representation of that word becomes more abstract due to the partial activation of all of the stored episodic traces. As such, an exemplar model would predict that each time a new episodic trace is generated it should refine the quality of the 'averaged' retrieved representation of that item (Burton, et al., 2005), making the representation of the phonological form of a novel word more robust to changes in talker or context. However, the effects of variability during study may depend on which aspects of the input are variable and which remain constant across different training tokens.

If talker information remains constant across different training instances of a novel word while speech rate and intonation differ (*within-talker variability*; Experiment 6) then all episodic traces will contain the same talker information. As a result, TSEs should still be observed when participants are later required to recognise the studied novel words. It is possible however that TSEs observed following within-talker variability training may be smaller than those observed in Experiments 1-3 since the TSEs in these earlier experiments may have resulted from a combination of many different types of specificity effects (talker identity, speech rate, intonation *etc.*). If this is the case then Experiment 6 should provide a better estimate of 'pure' TSEs for newly-learned words.

Introducing multiple talkers during the study phase of the experiment (*between-talker variability;* Experiment 7) should result in different episodic traces of a novel word containing different talker information. At retrieval, traces are assumed to be activated in accordance to their similarity to the input, allowing traces that match the test probe in terms of talker information to be activated more strongly than all of the other traces. However, since all episodic traces of a novel word contain the same phonological information as the test probe, all of these traces should be at least partially activated during recognition of the novel word. As a result, traces containing talker information that does not match the test probe will also contribute to the retrieved representation of a novel word, decreasing the size of the TSEs for items trained using between-talker variability compared to items trained in only one voice.

To summarise, an exemplar model of lexical representation would predict that TSEs for newly-learned words should decrease in size as more variability is introduced during study. However, it is important to note that since there are no pre-established episodic traces in memory for these novel nonwords prior to the study phase of the experiment, TSEs will be based on the same set of traces, all weighted equally (since they were all acquired at approximately the same point in time and should decay to roughly the same extent over time), at all test points. As such, an exemplar model would not predict that the size of the TSEs should change significantly over time for newly-learned words following either within-talker or between-talker variability during study.

Within a hybrid model variability during study should increase the number of episodic traces contained within the episodic subsystem, just as in an exemplar

model, resulting in smaller TSEs immediately after study following between-talker variability training, and also possibly following within-talker variability during study. However, variability may also affect the abstract subsystem since variability in the speech input may allow learners to determine which aspects of this input are invariant, and thus should promote generation of more abstract representations. Marsolek (2003) highlights three different types of information that are important in classifying items as belonging to the same category: (1) *presence-diagnostic features* are almost always present in items belonging to a particular category; (2) *absence diagnostic features* are almost always absent from items belonging to a particular category; and (3) *non-diagnostic features* may or may not be present in the input, but are not useful in defining the particular category in question. If listeners are given only a single token of a novel word (as in Experiments 1-5), they will be unable to differentiate between these three types of features; it is only given further instances of the same novel word that learners will be able to determine which features are diagnostic and which are non-diagnostic. Consistent with this suggestion Singh (2008) notes that increasing the variability in the input may lead the listener to classify the highly varying dimensions as irrelevant cues to lexical identity. Thus, as the amount of variability in the input during study increases, the abstract phonological representations that are assumed to be stabilised and strengthened over time in a hybrid model may be strengthened more rapidly. As such, these abstract representations would become more robust and should then contribute more to recognition of newly-learned words at delayed test points resulting in a decrease in the size of TSEs over time.

If talker identity varies across training tokens alongside variation in speech rate and intonation (*between-talker variability*) then only the phonological form of an item will be invariant across the different study tokens of a novel word, promoting the establishment of abstract phonological representations and resulting in a decrease in TSEs over time. However, if talker identity remains stable, but speech rate and intonation differ across tokens (*within-talker variability*), then participants may assume that speech rate and intonation are irrelevant aspects of the input, but that talker identity, as well as information about the phonological form of a novel word, is in fact relevant. If this is the case then TSEs should be maintained over time following within-talker variability during study since talker information will be deemed a diagnostic feature, and thus information about this feature should be

retained in memory. A further (but rather extreme) prediction is that TSEs may in fact become stronger over time as result of using within-talker variability during study if learners consolidate and strengthen information about all invariant properties of the input, in this case both the phonological information and the talker information. According to this more extreme prediction we might also expect that explicit identification of the training voice would improve, and the tentative pattern of talker-specific lexical competition effects observed in Experiments 2 and 3 should be strengthened.

The two experiments reported in this chapter explore whether increased variability in the speech input during training affects the retention of TSEs in recognition memory over the course of one week for novel words. They also explore whether talker variability during study affects the tentative pattern of talker-specific lexical competition effects observed in Experiments 2 and 3. Experiment 6 investigates the effects of within-talker variability whilst Experiment 7 explores the effects of between-talker variability. As in Experiments 2 and 3 participants studied a list of 24 novel nonwords before completing tests of both lexical competition and recognition memory at three different time points (Day 1, Day 2, and Day 8).

5.2 Experiment 6

During the study-phase of Experiment 6 participants heard half of the novel nonwords consistently spoken by a male talker, and half consistently spoken by a female talker, as in Experiments 1-3. However, rather than hearing a single token of each novel item repeated 18 times during the phoneme monitoring task, 18 unique tokens differing in speech rate, and as much as possible in intonation, were heard.

Both exemplar and hybrid models of lexical representation predict that TSEs should be smaller for newly-learned words following within-talker variability during study. Moreover, both types of model predict that TSEs should remain stable at delayed test points. Talker information associated with each item should also be explicitly accessible at all time-points, consistent with previous experiments. A hybrid model may also predict that TSEs might be strengthened following within-talker variability during study if the introduction of any type of variability in the input strengthens memory for the invariant properties of the input. If this latter prediction is correct then explicit recall of information about the study talker of each

novel nonword may improve, and the pattern of lexical competition effects associated with these novel items may become more strongly talker-specific.

5.2.1 Method

Participants

Thirty-two undergraduate students (*age range* = 18-21 years, 18 male) from the University of York completed the experiment. Ten additional participants were tested but were replaced due to failure to complete all three test sessions (4), experimenter error (1), having an error score more than 2.5*SD* above the mean in the phoneme monitoring study task (3) or reporting the test voice rather than the study voice in the male/female categorisation task (3).

Stimuli

The stimuli consisted of the same word triplets that were used in Experiments 1 to 3. However, all items were re-recorded for the current experiment. Eighteen tokens of each novel nonword were recorded for the phoneme monitoring task by the same male and female talkers as used in Experiments 1-3. The talkers were instructed to vary their intonation and speed of pronunciation for each novel nonword as much as possible whilst still producing natural-sounding tokens. An additional token of each novel nonword was recorded for the old-new categorisation task using an average speech rate and 'normal' intonation. Foil nonwords for the old-new categorisation task, as well as basewords and filler items for the lexical decision task were also re-recorded to avoid any potential differences in recording or voice quality between the current and previous recording sessions.

All stimuli were recorded in a sound attenuated booth using a Tascam DR-100 recorder and Sennheiser ME40 microphone. The stimuli were digitized at a 44.1 Hz sampling rate with 16-bit analogue-to-digital conversion, and peak amplitude was normalised using Adobe audition. On average, stimuli spoken by the male talker were slightly shorter than those spoken by the female talker. Paired samples t-tests indicated that this difference was significant for all groups of stimuli (study novel nonwords – $t(47) = 13.81$, $p < .001$; test novel nonwords – $t(47) = 15.78$, $p < .001$; foil nonwords – $t(47) = 16.03$, $p < .001$; basewords – $t(47) = 13.83$, $p < .001$).

Note that the counterbalancing of stimuli and talker for the phoneme monitoring and old-new categorisation tasks was identical to the counterbalancing

used in Experiment 2. Changes to the counterbalancing of stimuli in the lexical decision task are described below.

Design

Participants were tested individually in a sound-attenuated room. All equipment was the same as in previous experiments except that tasks were run on a Dell Vostro 230 computer. As in Experiment 3 each participant completed three sessions; one on Day 1, one on Day 2 approximately 24 hours later, and one on Day 8, one week after the first session. In the first session participants were familiarized with the novel words during the study phase of the experiment. Participants then complete the lexical decision task and old/new categorisation task (including confidence ratings and a male/female judgment) immediately after training. The session on Day 1 lasted approximately 45 min. In the experimental sessions on Days 2 and 8 participants completed only the lexical decision task and the old/new categorisation task. These latter two sessions lasted approximately 20 min each.

Procedure

The phoneme monitoring task was identical to that used in Experiments 2-3 except that 18 different tokens of each novel word were heard. Training tokens were ordered according to stimulus duration, and were split into three groups of six tokens – slow, medium, and fast. Within each of the six blocks of phoneme monitoring one slow, one medium, and one fast token of each novel word was heard. The order of these three tokens was randomised within each block.

In the *testing phase* of the experiment participants completed two tasks; one test of lexical competition (lexical decision), and one test of talker-specificity effects (old-new categorisation). The lexical decision task was identical to Experiment 2 except that half of the participants heard only the female talker during the lexical decision task, and half heard only the male talker (as in the pause detection task in Experiment 3). For all participants this manipulation resulted in half of the basewords being heard in the same voice that the corresponding novel nonword was training, and half being heard in a different voice, allowing us to examine whether any lexical competition effects observed were talker-specific or not. The old/new categorisation task included confidence ratings and male/female judgements, as in Experiments 3-5.

5.2.2 Results

Study phase

Sixteen participants were exposed to List 1 (11 male), and 16 to List 2 (7 male). Prior to analysis two items were removed from the data set due to a programming error which resulted in these items having greater than/less than 18 exposures during the study task. Once these items had been removed from the data set, the mean error rate in the phoneme monitoring task was 5.1% (*SD* = 2.3%), indicating that participants were paying close attention to the phonological form of the novel words during the study phase of the experiment. A repeated-measure ANOVA, with factors study talker (male *vs.* female), and list (1 *vs.* 2) showed that the main effect of list was marginally significant, $F_1(1,30) = 3.78$, $p = .061$, $\eta_p^2 = .11$, $F_2(1,44) = 3.51$, $p = .068$, $\eta_p^2 = .07$, with more errors overall for List 2 items (*M* = 5.9%, *SD* = 2.7%) compared to List 1 items (*M* = 4.2%, *SD* = 1.5%). However, there was no main effect of study talker, $F < 1$, nor was there a significant interaction between list and study talker, $F < 1$, $F_2(1,44) = 1.64$, *ns*. As such, whilst participants made more errors to List 2 items, this was not influenced by study talker, and thus is unlikely to have impacted on any subsequent TSEs in the test-phase of the experiment.

The mean RT in the phoneme monitoring task was 1310ms (*SD* = 251ms). A repeated-measures ANOVA, with the same factors as above, revealed that there was a significant main effect of study talker, $F_1(1,30) = 47.42$, $p < .001$, $\eta_p^2 = .62$, $F_2(1,44) = 10.66$, $p < .01$, $\eta_p^2 = .19$, with participants responding faster to items heard in the male voice (*M* = 1274ms, *SD* = 244 ms) than items heard in the female voice (*M* = 1345ms, *SD* = 261 ms). As in previous experiments, this main effect of study talker most likely reflects the differences observed in stimulus duration, with male items having shorter durations on average than female items. Thus, the relevant phonological information needed to make phoneme monitoring judgements would have become available sooner in the male tokens than in the female tokens. In the by-items analysis there was also a significant main effect of list, $F_2(1,44) = 26.55$, $p < .001$, $\eta_p^2 = .38$. However, this main effect was not significant in the by-participants analysis, $F_1(1,30) = 2.30$, *ns*, and the interaction between study talker and list was not significant in either analyses, $F < 1$, indicating that any advantage seen for items heard in the male voice at study was similar for both lists of items.

Talker-specificity effects

In the *old/new categorisation* task one participant, and three items produced error scores more than 2.5*SD* above the grand mean, and were removed prior to analysis. With these items removed participants responded correctly to 78.5% (*SD* = 5.9%) of the items when making old/new categorisation decisions. RTs were measured from word onset until participants made an old/new categorisation button-press response (*M* = 2096ms, *SD* = 314ms).

Analysis of SDT data using a repeated-measures ANOVA with factors test-phase talker (same *vs.* different), and day (1,2, and 8) revealed that for d′ values (Figure 5.1a) there was a significant main effect of test-phase talker, $F(1,29) = 10.62$, $p < .01$, $\eta_p^2 = .27$, but a non-significant main effect of day $F(2,58) = 1.10$, *ns*, and no interaction between test-phase talker and day, $F < 1$. In order to determine whether there was a significant same-talker advantage at all time points the data were analysed separately for each test session. Analysis showed that the main effect of test-phase talker was significant on Day 1, $F_1(1,29) = 5.04$, $p < .05$, $\eta_p^2 = .15$, but only marginally significant on Day 2, $F_1(1,29) = 3.11$, $p = .088$, $\eta_p^2 = .10$, and Day 8, $F_1(1,29) = 3.77$, $p = .06$, $\eta_p^2 = .12$. Nonetheless, as noted above, neither the main effect of day, nor the interaction between test-phase talker and day were significant suggesting that the size of the TSEs did not change significantly over time.

For β values (Figure 5.1b) there was also a main effect of test-phase talker, $F_1(1,29) = 46.42$, $p < .001$, $\eta_p^2 = .62$, a non-significant main effect of day, $F < 1$, and a non-significant interaction between test-phase talker and day, $F < 1$, indicating that the differences in bias for same- and different-talker items did not change over the course of a week. Further analysis confirmed that TSEs were significant at all time points individually in the β data (Day 1 – $F_1(1,29) = 24.59$, $p < .001$, $\eta_p^2 = .46$; Day 2 – $F_1(1,29) = 19.13$, $p < .001$, $\eta_p^2 = .40$; Day 8 – $F_1(1,29) = 13.97$, $p = .001$, $\eta_p^2 = .33$).

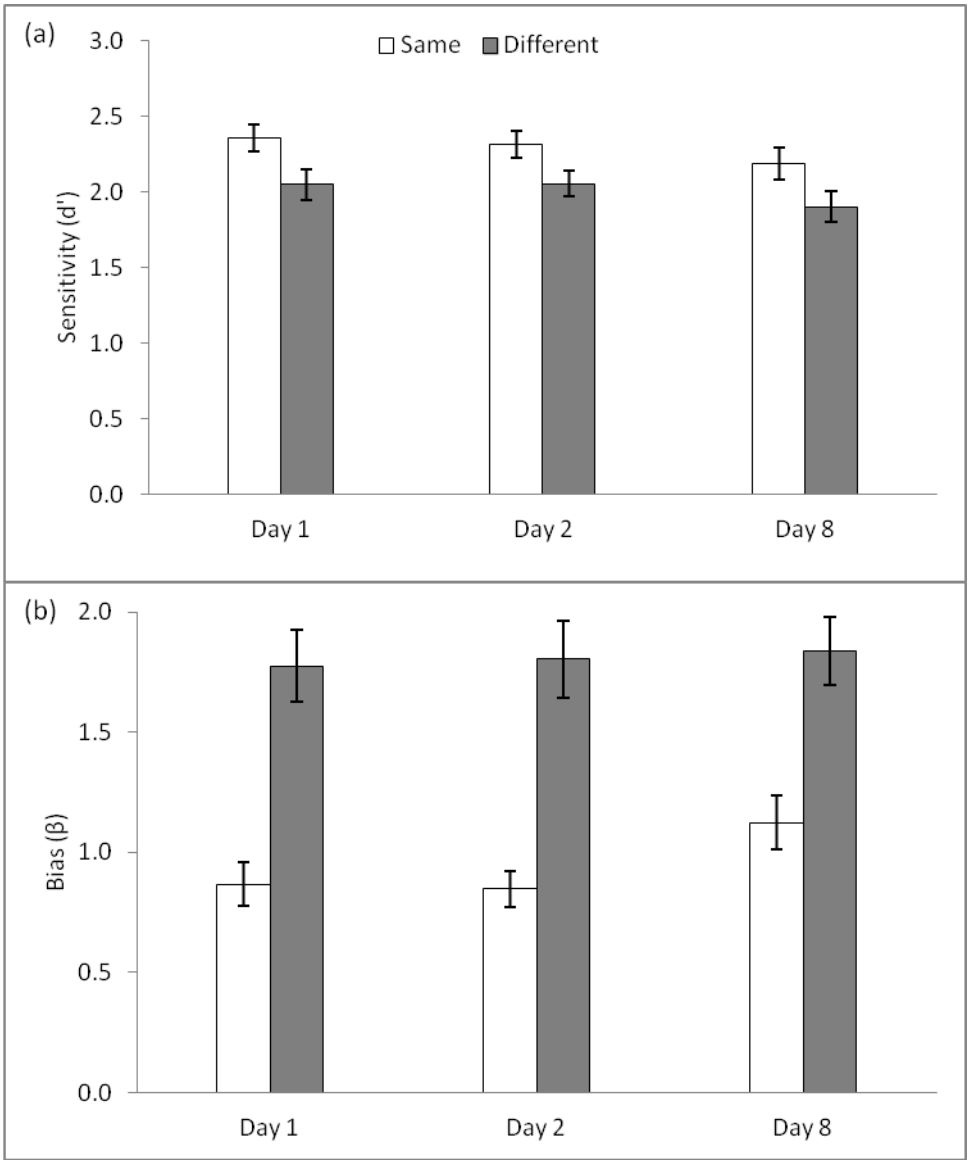

Figure 5.1. (a) Sensitivity and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Data from the *confidence ratings* were plotted as ROC curves and AUC values were calculated for same- and different-talker items in each test session as in Experiments 3-5. Analysis of AUC values (Figure 5.2) using a repeated-measures ANOVA with factors test-phase talker (same *vs.* different) and day (1, 2, *vs.* 8) revealed significant main effects of both test-phase talker, $F(1,29) = 5.34$, $p < .05$, $\eta_p^2 = .16$, and day, $F(1.6, 46.4) = 3.74$, $p < .05$, $\eta_p^2 = .11$, indicating that AUC values were greater for same-talker items, but that overall they decreased across the three test sessions. The interaction between test-phase talker and day was non-significant,

F < 1, suggesting that the size of the TSEs did not differ significantly over time. However, post-hoc analyses revealed that the main effect of test-phase talker was significant only on Day 2, $F(1,29) = 5.00$, $p < .05$, $\eta_p^2 = .15$, although it was marginally significant on Day 8 also, $F(1,29) = 3.21$, $p = .084$, $\eta_p^2 = .10$.

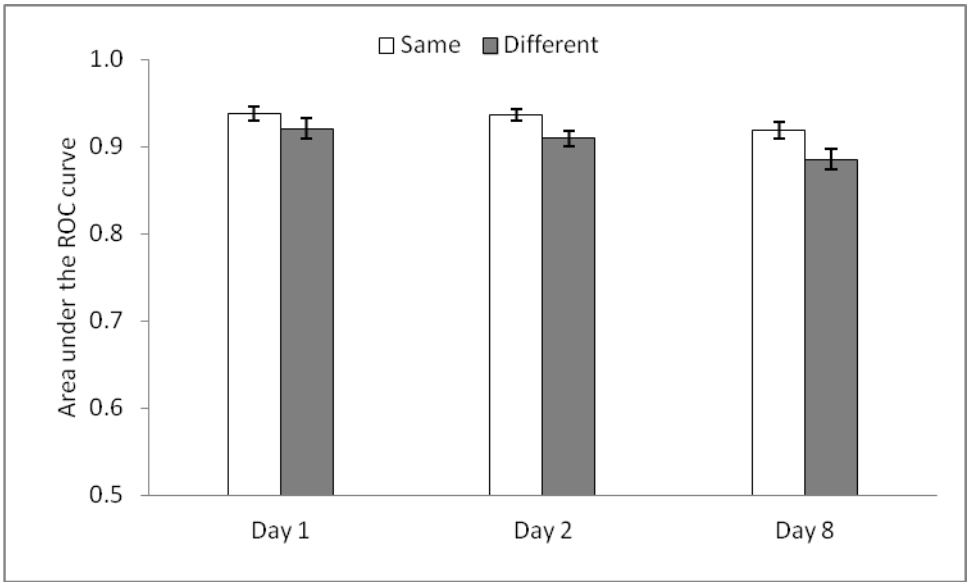

Figure 5.2. Confidence ratings given to old/new categorisation decisions were used to plot ROC curves. Area under the ROC curve was calculated separately for same- and different-talker items in each test session. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Mean accuracy in the *male/female categorization* task was 69.1% (*SD* = 9.6%), which was significantly above chance, $t(31) = 11.34$, $p < .001$. Analysis of the data (Table 5.1) revealed that there was a significant main effect of test-phase talker, $F_1(1,30) = 32.70$, $p < .001$, $\eta_p^2 = .52$, $F_2(1,44) = 128.30$, $p < .001$, $\eta_p^2 = .75$, with participants responding more accurately to same-talker compared to different-talker items. There was also a significant main effect of day, $F_1(2,60) = 3.37$, $p < .05$, $\eta_p^2 = .10$, $F_2(2,88) = 4.32$, $p < .05$, $\eta_p^2 = .09$. The interaction between test-phase talker and day was also significant by-participants, $F_1(2,60) = 3.51$, $p < .05$, $\eta_p^2 = .11$ $F_2(2,88) = 1.74$, *ns*, although the same-talker advantage was significant at all time points individually (Day 1 - $F_1(1,30) = 30.83$, $p < .001$, $\eta_p^2 = .51$, $F_2(1,44) = 92.85$, $p < .001$, $\eta_p^2 = .68$; Day 2 - $F_1(1,30) = 43.94$, $p < .001$, $\eta_p^2 = .59$, $F_2(1,44) = 88.89$, $p <$

.001, η_p^2 = .67; Day 8 - $F_1(1,30) = 12.02$, $p < .01$, η_p^2 = .29, $F_2(1,44) = 28.03$, $p < .001$, η_p^2 = .39.

Table 5.1. Percentage of correct responses in the male/female categorisation task (novel nonwords only), split according to whether the item was spoken in the same or a different talker to study.

Day	Same	Different
1	84.5	57.2
2	83.4	51.4
8	76.4	54.1

Lexical competition effects

In the lexical decision task participants performed accurately across all items, with a mean accuracy score of 92.9% (*SD* = 5.9%). Data from the 48 basewords were filtered using the same criteria as used for the lexical decision data in Experiment 2 resulting in 8.3% of data points being removed from the baseword data set prior to analysis. The mean RT was 991ms (*SD* = 95ms). Two participants had error scores more than 2.5*SD* above the grand mean; data from these two participants were removed prior to analysis. A repeated-measures ANOVA, with factors word-type (test *vs.* control), and day (1, 2, or 8) revealed that there was a significant main effect of day, $F_1(2,56) = 11.32$, $p < .001$, η_p^2 = .29, $F_2(2,88) = 32.18$, $p < .001$, η_p^2 = .42, with RTs decreasing across session, most likely due to task repetition/practice effects. The main effect of word-type was non-significant, $F_1(1,28) = 2.89$, *ns*, $F_2(1,44) = 1.21$, *ns*, as was the interaction between word-type and day, $F_1(2,56) = 2.39$, *ns*, $F_2(2,88) = 1.68$, *ns*.

Figure 5.3a indicates that participants showed an unusual pattern of lexical competition, with an increase in lexical competition between Days 1 and 2, but a decrease in lexical competition between Days 2 and 8. Analysis of the Day 2 data on its own showed that there was a main effect of word-type in this session, $F_1(1,28) = 6.46$, $p < .05$, η_p^2 = .19, $F_2(1,44) = 5.00$, $p < .05$, η_p^2 = .10, with slower RTs to test basewords than control basewords, suggesting that lexical competition effects were significant in this test session. However, 13 out of the 30 participants included in the analysis showed the unusual pattern of lexical competition effects described above. Further analyses confirmed that this unusual pattern of data did not depend on whether participants heard the male or the female voice during the lexical decision

task, nor did it depend on the speed of responding when a median split was used to divide participants into fast versus slow responders. Thus it remains unclear why lexical competition effects emerged on Day 2 but then disappeared on Day 8.

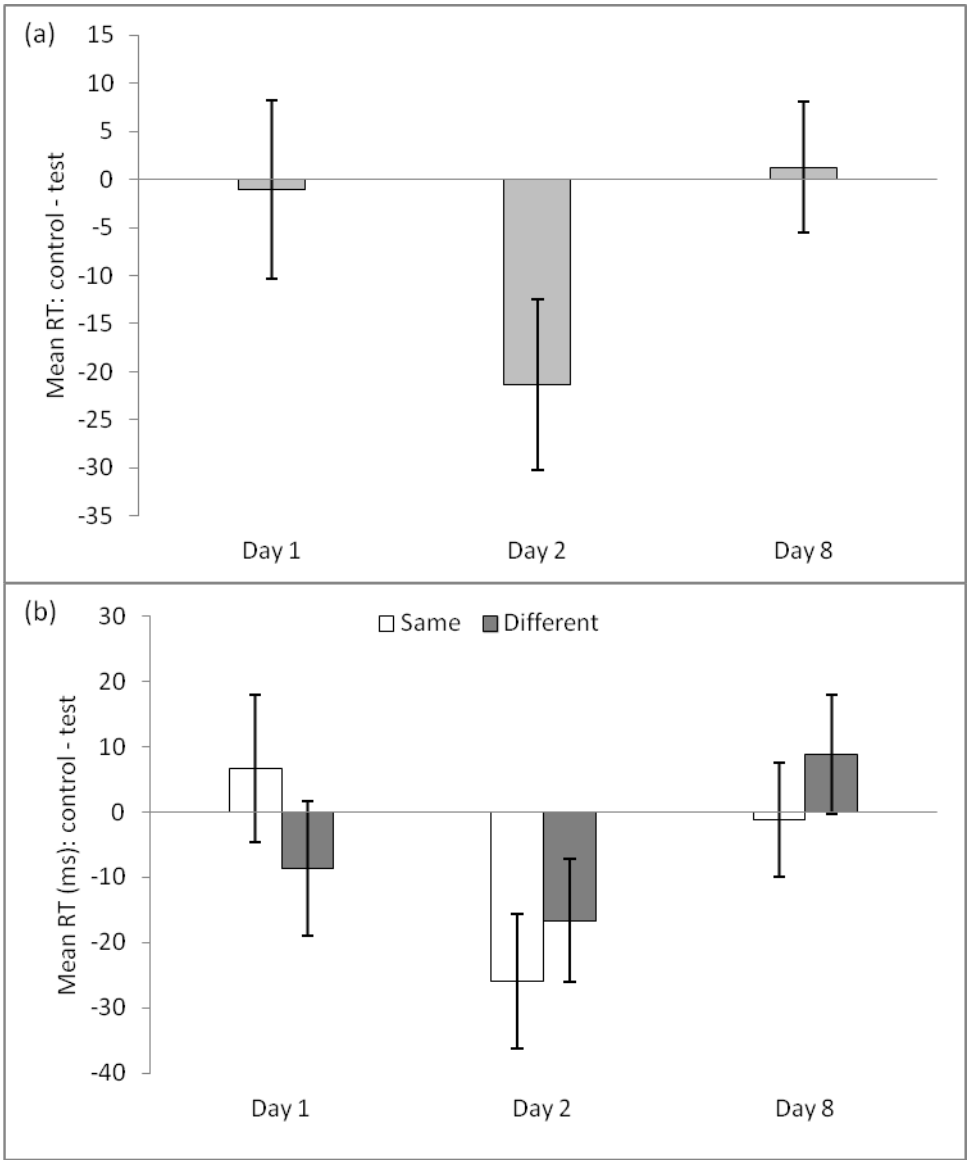

Figure 5.3. (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) basewords in the lexical decision task. (b) Lexical decision data split according to whether the test baseword was spoken in either the same voice that the corresponding novel word was trained in, or a different voice. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

One possibility is that RTs in the lexical decision task may have been influenced by whether the baseword was heard in a same or different talker to that in which the novel nonword was studied, as suggested by the patterns of data observed in Experiments 2 and 3. To investigate this possibility an additional set of analyses were carried out (Figure 5.3b) using a repeated-measures ANOVA, with factors day (1, 2, or 8) and baseword type (same-talker, different-talker, or control). However, this analysis revealed that only the main effect of day was significant, $F_1(2,56) = 9.94$, $p < .001$, $\eta_p^2 = .26$, $F_2(2,88) = 21.97$, $p < .001$, $\eta_p^2 = .33$, once again reflecting the speeding up of response times across sessions. As in the main analysis, the main effect of baseword type as non-significant, $F < 1$, as was the interaction between baseword type and day, $F_1(4,112) = 1.72$, *ns*, $F_2(4,176) = 1.98$, *ns*. Interestingly, post-hoc comparisons on the Day 2 data revealed that only RTs to same-talker items differed significantly from RTs to control items, $F_1(1,28) = 6.16$, $p < .05$, $\eta_p^2 = .18$, $F_2(1,44) = 3.22$, $p = .08$, $\eta_p^2 = .07$. The difference in RTs between different-talker items and control items was not significant on Day 2, $F_1(1,28) = 2.85$, *ns*, $F_2(1,44) = 2.78$, *ns*.

Correlations between talker-specificity and lexical competition effects

Correlations between lexical competition data and d′, β, and RT data from the old/new categorisation task were calculated as in Experiments 2 and 3. Analyses revealed that the change in β values between days 1 and 2 correlated with the change in lexical competition values between these days, $r = -.51$, $p = .004$, indicating that as lexical competition increased the difference between same- and different-talker biases decreased. No other correlations survived a Bonferroni correction.

5.2.3 Discussion

Experiment 6 demonstrates that, as in Experiments 1-5, highly-detailed representations of novel nonwords were generated during the study-phase of the experiment, and that these representations were capable of supporting significant TSEs in the d' data immediately after study. However, contrary to Experiments 2 and 3 TSEs in the d' data were only marginally significant on Days 2 and 8 after within-talker variability was included during the study phase of the experiment. Likewise, TSEs in the AUC data were less reliable in Experiment 6, reaching statistical

significance only on Day 2 (although numerically there was a trend towards a same-talker advantage at all time points); in Experiments 2 and 3 AUC analysis revealed significant TSEs in all three test sessions. Nonetheless, it is important to note that test-phase talker (same *vs.* different) did not interact with day in either the d' or AUC analyses in Experiment 6, suggesting that the size of TSEs did not change significantly over the course of a week. This finding is consistent with data from Experiments 2 and 3, and suggests that introducing within-talker variability did not change the pattern of TSEs over time even though the effects appear to be somewhat weaker overall compared to experiments in which there was no talker variability during study. One possible explanation for the fact that TSEs appear to be less reliable overall following within-talker variability during study is that the TSEs observed for novel nonwords in Experiments 1 to 5 may have resulted from a combination of a number of different types of specificity (*e.g.,* talker identity + speech rate + intonation). If this is the case then Experiment 6 provides a truer measure of TSEs during the recognition of newly-learned words.

Despite the stability of TSEs across the course of a week in the d' and AUC data accuracy in the male/female categorisation task decreased significantly over time, suggesting that the ability to explicitly access highly-detailed episodic representations decreased over the course of a week. An exemplar model of lexical representation may attempt to explain the decrease in performance in the male/female categorisation task by assuming that episodic traces decay over time, resulting in the loss of specific details about the study talker. A hybrid model may also assume that episodic representations decay gradually over time in the episodic subsystem, but could also explain the decrease in male/female categorisation performance in terms of a gradual change in reliance between episodic and abstract systems.

Interestingly, β data revealed significant TSEs in all test sessions, with no significant interaction between test-phase talker and day (unlike Experiments 2 and 3 in which this interaction was significant). This finding suggests that including within-talker variability during study leads to a more robust and long-lasting same-talker bias in the response criterion applied to the old/new categorization task. Thus, this aspect of our data supports the more extreme prediction made by a hybrid model that variability in the input serves to highlight invariant features (in this case both talker information and phonological information) and results in strengthening of

memory for these invariant features. However, it is interesting that it is only the β data that seem to show any kind of increased stability in TSEs as a result of within-talker variability during study.

In contrast to the presence of TSEs immediately after study in the current experiment, lexical competition effects did not emerge until Day 2, consistent with Experiment 2, the numerical trend observed in the pause detection data in Experiment 3, and a number of previous studies demonstrating the emergence of lexical competition effects between novel and existing words only after a period of sleep-associated offline consolidation (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Henderson et al., submitted-a, submitted-b; Tamminen & Gaskell, 2008). Interestingly Experiment 6 suggests that only same-talker items show significant lexical competition effects on Day 2, consistent with the numerical trend observed in Experiment 3. However, this finding contrasts with the pattern of lexical competition data observed in Experiment 2 which suggested that on Day 2 only between-talker items showed significant competition effects. Nonetheless, it is important to note that in the current experiment the lexical competition effects were absent on Day 8, a finding that is inconsistent with Experiment 2, as well as a study by Tamminen and Gaskell (2008) demonstrating that lexical competition effects associated with newly-learned pseudowords were observable up to 8 months after novel nonwords had initially been studied, suggesting that these effects were relatively stable and long-lasting and as such, were likely to be due to the establishment of robust new lexical representations. Thus, it would be unwise to place too much emphasis on the different patterns of talker-specificity lexical competition effects observed between Experiments 2 and 6 as a result of the absence of lexical competition effects on Day 8 in Experiment 6.

5.3 Experiment 7

Between-talker variability was introduced during the study-phase of Experiment 7 such that each novel nonword was encountered in both a male and a female voice during the phoneme monitoring task. Half of the items were consistently spoken by a combination of Male 1/Female 1, and the other half were spoken consistently by a combination of Male 2/Female 2. As in previous

experiments, at test half of the items were heard in one of the same voices as study whilst the other half were heard in a different voice.

As outlined in the introduction to this chapter both exemplar and hybrid models of lexical representation predict that between-talker variability during training should result in smaller TSEs compared to Experiments 1-6 where either no-variability or within-talker variability was encountered during study. However, exemplar models predict that the size of the TSEs should remain constant over time whilst hybrid models predict that TSEs should decrease in size over time since increased variability in the input may promote the development and stabilisation of abstract representations, speeding up the change in reliance between the episodic and abstract subsystems.

5.3.1 Method

Participants

Thirty-two undergraduate students (*age range* = 18-23 years, 10 male) from the University of York completed the experiment. Eight additional participants were tested but were replaced due to failure to complete all three test sessions (5), technical failure (2), or having an error score more than 2.5 *SD* above the mean in the phoneme monitoring task (1).

Stimuli

The stimuli consisted of the same word triplets that were used in Experiment 6. For the purpose of the current experiment, the two talkers used in Experiment 6 will be referred to as Male 1 and Female 2. For the phoneme monitoring task nine tokens of each novel word were selected from the tokens recorded by Male 1 and Female 2 for the previous experiment. Two additional speakers, one male (Male 2) and one female (Female 1), also recorded nine tokens of each novel word, with variations in intonation and speed of pronunciation. Stimuli were recorded and edited in the same manner as those in the previous experiment. The audio files used in the test phase of the experiment were identical to those used in Experiment 6.

Mean stimulus duration was calculated across the 9 tokens of each novel nonword heard in phoneme monitoring for each talker. Paired-samples t-tests revealed that although there was no difference in mean stimulus duration for the two female talkers, $t(47) = -.68$, *ns*, Male 2 produced the items that were significantly

faster than Male 1, $t(47) = 7.41$, $p < .001$. Moreover, when mean stimulus duration was calculated for Male 1/Female 1 combined, and compared to Male 2/Female 2 combined there was a significant difference, $t(47) = 5.26$, $p < .001$, with items spoken by the combination of Male 2/Female 2 having faster speech rates on average

Design and Procedure

The phoneme monitoring task was altered in the current experiment such that each novel word was heard in two voices, one male and one female, during the study-phase of the experiment. In order to allow the same test materials to be used, the male speaker used in Experiment 6 (Male 1) was paired with the new female speaker (Female 1), and the female speaker that had been used in Experiment 6 (Female 2) was paired with the new male speaker (Male 2). Hence, half of the novel words in the phoneme monitoring task were spoken consistently by Male 1 and Female 1, whilst the other half were spoken consistently by Male 2 and Female 2. Items were encountered 18 times during the phoneme monitoring task, with 9 tokens spoken by each of the talkers. As in Experiment 6, tokens for each talker were ordered according to stimulus duration and were split into three groups of tokens – slow, medium, and fast. Within each of the six blocks of phoneme monitoring one slow, one medium, and one fast token of each novel word was heard in a random order. All four speakers were included in each block of phoneme monitoring such that within each pair of voices, two tokens occurred in one of the voices, and one in the other (e.g., 2 female tokens, and 1 male token), with the number of tokens per talker alternating between blocks such that if 2 female tokens and 1 male token were heard in the first block, then 2 male tokens and 1 female token would be heard in the second block for the same item, and so on.

The test-phase of the experiment was identical to Experiment 6 except that the source memory judgment, in which participants were asked to indicate whether the item had originally been studied in either a male or a female voice, was omitted due to the fact that all novel words were encountered in both a male and a female voice during the phoneme monitoring task. As in Experiment 6, participants completed the test-phase of the experiment at three time points; immediately after study (Day 1), one day later (Day 2), and one week later (Day 8).

5.3.2 Results

Study phase

Sixteen participants were exposed to List 1 (6 male) and sixteen to List 2 (4 male). The mean error rate in the phoneme monitoring task was 5.4% (*SD* = 2.1%), indicating that participants were paying close attention to the phonological form of the novel words during the study phase of the experiment. A repeated-measures ANOVA, with factors study-phase talker (male 1/female 1 *vs.* male 2/female 2), and list (1 *vs.* 2) revealed non-significant main effects of study-phase talker, $F < 1$, and list, $F_1(1,30) = 1.69$, *ns*, $F_2(1,46) = 1.37$, *ns*, and a non-significant interaction between study-phase talker and list, $F_1(1,30) = 2.88$, *ns*, $F_2(1,46) = 1.20$, *ns*.

The mean RT in the phoneme monitoring task was 1270ms (*SD* = 299ms). Analysis of RTs revealed non-significant main effects of study talker, $F < 1$, and list, $F_1(1,30) = 1.10$, *ns*, although the main effect of list was significant in the by-items analysis, $F_2(1,46) = 19.43$, $p < .001$, $\eta_p^2 = .30$, reflecting faster responses to List 1 items (*M* = 1224ms, *SD* = 258ms) compared to List 2 (*M* = 1331ms, *SD* = 328ms). Nevertheless the interaction between study-phase talker and list was not significant, $F < 1$.

Talker-specificity effects

In the *old/new categorisation* task two items produced error scores more than 2.5*SD* above the grand mean and were removed prior to analysis. With these items removed participants responded correctly to 82.3% (*SD* = 6.7%) of the items when making old/new categorisation decisions. Mean RT, measured from word onset until participants made an old/new categorisation button-press response, was 1767ms (*SD* = 293ms). For two participants old/new categorisation data from one of the three test sessions were lost due to a technical error. Data from the remaining two test sessions for these participants were included in the analyses.

A repeated-measures ANOVA with factors test-phase talker (same *vs.* different) and day (1,2, and 8) showed that for d′ values (Figure 5.4a) there was a significant main effect of day, $F(2,56) = 3.76$, $p < .05$, $\eta_p^2 = .12$. The main effect of test-phase talker was also marginally significant, $F(1,28) = 3.10$, $p = .089$, $\eta_p^2 = .10$. There was however no interaction between test-phase talker and day, $F(2,56) = 1.66$, *ns*. In order to determine whether the main effect of test-phase talker was significant

at any time point the data were analysed separately for each test session. Analysis showed that the main effect of test-phase talker was significant on Day 1, $F(1,29) = 8.09$, $p < .01$, $\eta_p^2 = .22$, but non-significant on Days 2, $F < 1$, and 8, $F < 1$.

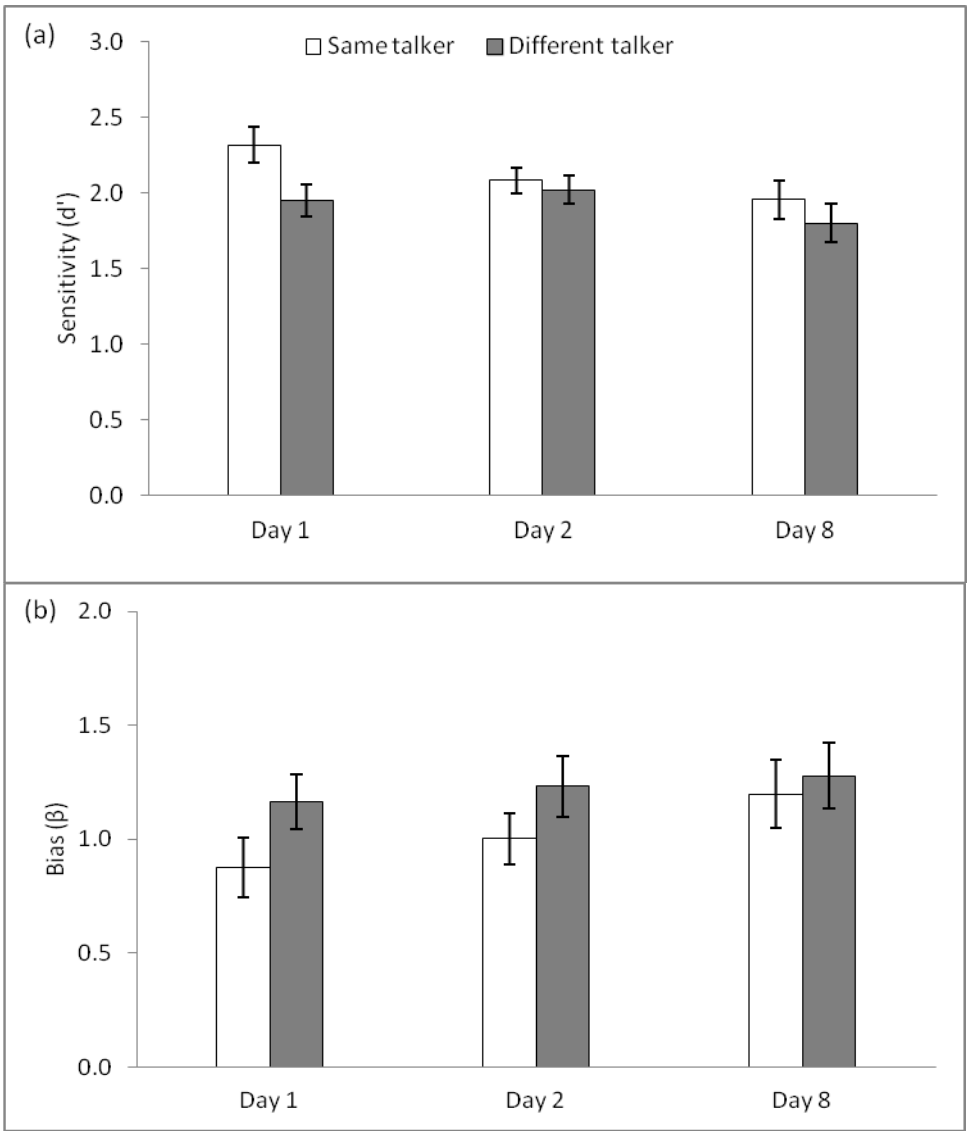

Figure 5.4. (a) Accuracy and (b) bias in the old/new categorisation task as a function of whether the study and test talkers were the same or different. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

For β values (Figure 5.4b) the main effect of test-phase talker was not significant, $F(1,28) = 2.06$, *ns*. The main effect of day was also non-significant, $F(2,56) = 1.15$, *ns*, as was the interaction between test-phase talker and day, $F < 1$ (Figure 5.4b), indicating that there were no differences in bias for same- and different-talker items at any time point. In order to verify this finding the data were

analysed further, broken down by test session. The main effect of test-phase talker did not approach significance in any of the test-sessions individually.

Data from the *confidence ratings* were plotted as ROC curves and AUC values were calculated for same- and different-talker items in each test session (Figure 5.5). One participant had a mean AUC score more than 2.5 *SD* below the grand mean. Data from this participant were removed from the data set prior to analysis. A repeated-measures ANOVA, with AUC values included as the dependent variable and factors day (1, 2, *vs.* 8), and test-phase talker (same *vs.* different) revealed that there was a significant main effect of day, $F(2,54) = 4.32$, $p < .05$, $\eta_p^2 = .14$, but no main effect of test-phase talker, $F(1,27) = 2.03$, *ns*, and no interaction between test-phase talker and day, $F < 1$. The main effect of test-phase talker was not significant at any of the test points when data from each were analysed individually.

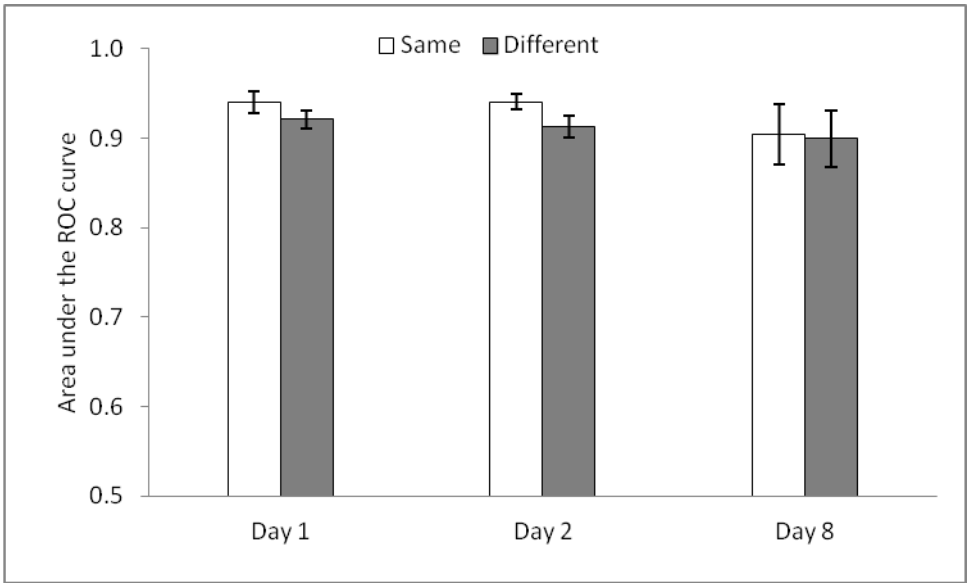

Figure 5.5. Confidence ratings given to old/new categorisation decisions were used to plot ROC curves. Area under the ROC curve was calculated separately for same and different talker items in each test session. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Lexical competition effects

Overall participants responded correctly to 91.3% (*SD* = 5.0%) of the items in the lexical decision task. Data from the 48 basewords were filtered as described in Experiment 2. Two participants (one exposed to each list) had mean RTs more than 2.5SD above the grand mean and were removed from the data set prior to analysis.

Likewise, two items (one from each list) had error scores more than 2.5*SD* above the grand mean in the by-items analysis and were removed. Overall 7.8% of data points were removed from the baseword data set prior to analysis. A repeated-measures ANOVA, with factors day (1, 2, *vs.* 8) and baseword type (test *vs.* control) revealed a marginal main effect of day, $F_1(2,52) = 2.79$, $p = .07$, $\eta_p^2 = 10$, $F_2(2,88) = 12.40$, $p < .001$, $\eta_p^2 = .22$, indicating, unsurprisingly, that RTs decreased over the three test session. The main effect of baseword type was non-significant by-participants, $F < 1$ (although it was marginally significant by-items, $F_2(1,44) = 3.32$, $p = .075$, $\eta_p^2 = .07$), as was the interaction between baseword type and day, $F < 1$ (Figure 5.6a). Separate analysis of data from each test session confirmed these findings, revealing non-significant main effects of baseword type at all time points (Day 1 – $F < 1$; Day 2 – $F_1 < 1$, $F_2(1,44) = 1.77$, *ns*; Day 8 – $F_1(1,26) = 1.07$, *ns*, $F_2(1,44) = 2.39$, *ns*), indicating that RTs did not differ between basewords with and without novel competitors at any time point.

As in Experiments 2, 3, and 6, additional analyses were conducted in order to determine whether lexical competition effects were dependent on whether the test basewords were heard in the same or a different voice as that in which the phonologically-similar novel nonwords were trained (Figure 5.6b). A repeated-measure ANOVA, including the variables day (1, 2, 8) and baseword type (same-talker, different-talker, control) revealed a main effect of day, $F_1(2,50) = 3.99$, $p < .05$, $\eta_p^2 = .14$, $F_2(2,88) = 12.21$, $p < .001$, $\eta_p^2 = .22$, as in the main analysis reported above. There was however no main effect of baseword type, $F_1 < 1$, $F_2(1.7, 76.2) = 2.55$, $p = .083$, $\eta_p^2 = .06$, or any interaction between day and baseword type, $F < 1$. Post hoc analysis of each test session individually did revealed a significant difference between same-talker and control items on Day 8 in the by-items analysis, $F_2(1,44) = 5.97$, $p < .05$, $\eta_p^2 = .12$. This effect was also marginally significant by participants, $F_1(1,26) = 2.74$, $p = .11$, $\eta_p^2 = .10$. However the lack of significance by-participants questions the reliability of this finding. As such, whilst numerically the pattern of data in the current experiment is very similar to that observed in Experiment 2 (Figure 3.2b), these effects were not as robust.

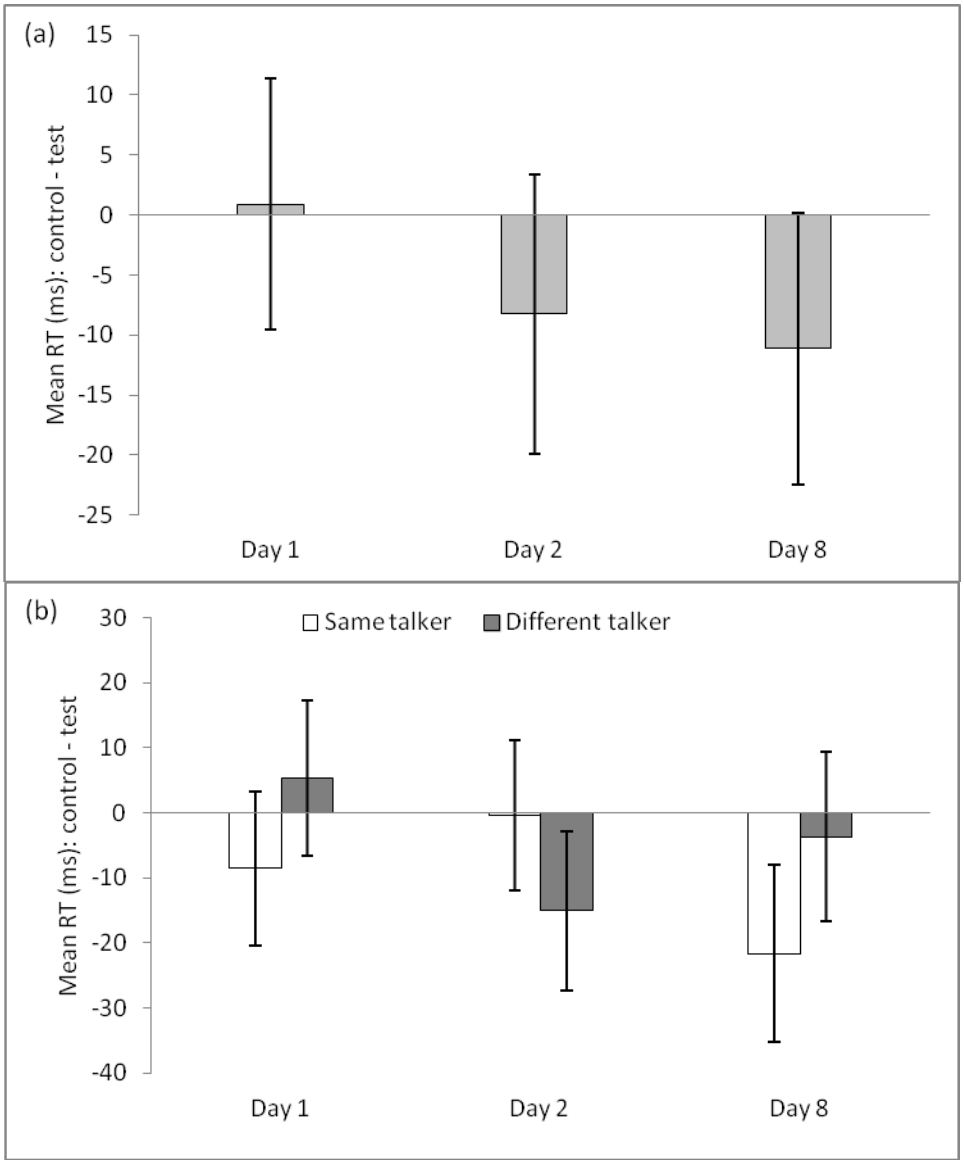

Figure 5.6. (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) basewords in the lexical decision task. (b) Lexical decision data split according to whether the test baseword was spoken in the same voice that the corresponding novel word was trained in, or a different voice. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Correlations between talker-specificity and lexical competition effects

Correlations between lexical competition effects and TSEs were calculated in the same way as in previous experiments. None of the correlations survived a Bonferroni correction.

5.3.3 Discussion

Experiment 7 demonstrates that even when novel nonwords are heard in more than one voice during study participants are still able to form robust phonological representations of those words and show accurate recognition of those items when tested up to a week later. Notably, Experiment 7 does show a significant decrease in overall old/new categorisation performance over the course of a week. A similar decrease in overall accuracy was not observed following either no variability (Experiments 3) or within-talker variability (Experiments 6) during study (although the main effect of day was marginally significant in Experiment 2 ($p = .064$)). This finding indicates that the representations supporting old/new recognition decisions may decay at a slightly faster rate after novel words have been heard in more than one voice compared to when they are studied in only a single voice.

Despite this decrease in performance over time, overall d′ scores still revealed a (non-significant) trend towards TSEs in the old/new categorisation task. As in Experiments 1-6 TSEs were statistically significant in the d′ data when participants were tested immediately after studying the novel nonwords, indicating that information about each of the two study talkers for each novel nonword must have been encoded and stored in memory alongside the phonological form of a word. This finding is important for two reasons. Firstly, the presence of significant TSEs on Day 1 (and the marginal significance of TSEs overall) argues against the suggestion that TSEs were present in Experiments 1-6 only because two talkers of different genders were used (*voice connotation hypothesis*, Geiselman & Crawley, 1983). In Experiment 7 participants heard all novel nonwords spoken by one male and one female talker during study. Thus, the presence of TSEs in the immediate old/new categorisation test must depend upon retention of specific details about each study talker, not simply the presence of different gender tags associated with each novel nonword (consistent with Goldinger, 1996, and Palmeri et al., 1993).

Secondly, is important to consider the fact that there were only nine tokens of each novel nonword associated with each of the two study voices in Experiment 7. In Experiments 1, 2, 3, and 6 there were 18 tokens of each novel nonword associated with the talker used during study. Therefore, it could be argued that the non-significant TSEs on Days 2 and 8 in the d′ data arose due to less robust representation of talker information in memory. However, if this were the case then

TSEs should also have been absent on Day 1. This was not the case in the d' analysis although notably TSEs were absent at all time points in the AUC data in this experiment. Interestingly, TSEs were also non-significant in the β data in all test sessions suggesting that talker identity was not used as a cue to whether an item was old or new. The pattern of β data contrasts with that observed in Experiment 6 where the same-talker response bias in the old/new categorisation task was significant and stable over time following within-talker variability during study. This latter finding suggests that participants are able to alter their response biases depending on which information they deem to be important based on the structure and variability of the input during training.

The time-course of TSEs in the d' data is of particular interest. Specifically, the pattern of TSEs over the course of a week following between-talker variability during study was very similar to that observed in Experiment 5 for existing words, with significant TSEs in the immediate test session but non-significant TSEs in the delayed test sessions. Whilst the non-significant interaction between test-phase talker and day in the current experiment, as well as for existing words in Experiment 5, makes it difficult to claim that the time-course of TSEs is *significantly* different to that observed for novel nonwords studied in a single voice, the similarity between existing words and novel words studied in more than one voice is intriguing.

In Experiment 7, an exemplar model would assume that immediately after study there should be (according to extreme models such as MINERVA 2; Goldinger, 1998) 18 unique traces of each novel nonword, one associated with each of the unique study tokens (nine containing information about the male study talker and nine containing information about the female study talker). All of these traces should be weighted equally since they were acquired at approximately the same time. Thus, the presence of TSEs for novel nonwords immediately after study in Experiment 7 must be accounted for by activation of each of these episodic traces in accordance to their similarity to the speech input at test. Over time all traces may decay to some extent, and some details may be lost as a result of this trace decay. However, in simulations using MINERVA 2 the forgetting rate is typically fixed (Goldinger, 1998), and as such it may be assumed that the amount of decay is roughly equivalent for all 18 traces. Thus when participants are re-tested on Days 2 and 8 traces should again all be weighted equally initially, but activated in accordance to their similarity to the speech input at test. As a result, an exemplar

model of lexical representation would predict that TSEs should be a similar size at all test points since. This prediction is consistent with the absence of a significant interaction between test-phase talker and day in the d′ data from Experiment 7.

In a hybrid model, immediately after study, recognition of newly-learned words is assumed to rely on the episodic subsystem, as in exemplar models. However, over time abstract representations of the novel nonwords may become established in the abstract subsystem and begin to play a greater role in recognition and processing of newly-learned words. Providing training tokens that vary in talker identity, speech rate and intonation may allow the listener to home-in on the invariant properties of the speech input, namely the phonological form of the novel items, and to form robust abstract representations of these items more quickly than is possible when no variability is included during training (Experiment 1-5), or indeed when only within-talker variability is included during training (Experiment 6). However, if variability affects the robustness of abstract representations then larger and more robust lexical competition effects should have been observed in Experiment 7. This was not the case; lexical competition effects were non-significant at all time points. An alternative may be that it is not the strength of abstract representations that is affected by variability in the input, but rather the speed with which there is a change in reliance between the episodic and abstract subsystems in a hybrid model. Tentative evidence supporting this suggestion comes from the finding that overall performance in the old/new categorisation task, which is assumed to rely on recollection of highly-detailed episodic traces (Holdstock et al., 2002) decreased over time in Experiment 7. Likewise the non-significant TSEs in the d′ data on Days 2 and 8 point towards an increased reliance on abstract representations (in conjunction with episodic representations) in the current experiment. It is clear however that further experiments are required in order to disentangle the predictions of episodic and hybrid models. Perhaps including a greater number of talkers during study would provide more robust evidence regarding the effects of between-talker variability on the time-course of TSEs for newly-learned words that would allow us to differentiate between the predictions of episodic and hybrid models at delayed time points since the current data are unable to do so.

It is important to note that lexical competition effects were not significant at any time point in Experiment 7. There were however numerical trends towards

lexical competition on Days 2 and 8, consistent with Experiments 2 and 3, Day 2 of Experiment 6, and a number of previous studies examining lexical competition effect following word learning in adults (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008). Most interestingly the pattern of talker-specific lexical competition effects observed in Experiment 7 is remarkably similar to that observed in Experiment 2, with different-talker items showing numerically larger competition effects on Day 2 but same-talker items showing numerically larger competition effects on Day 8. It is possible that the lack of significant lexical competition effects across Experiments 3, 6, and 7 stems from a lack of power in the analyses since, if lexical competition effects are in fact talker-specific then our estimates of lexical competition are based on RTs to only 12 basewords with novel nonword competitors. Indeed, in the between-talker variability experiment each novel nonword was heard only nine times in each voice during study, further limiting the possibility of observing significant lexical competition effects if these are in fact talker-specific. Below we present a combined analysis of lexical decision data from Experiments 2, 6, and 7 in order to address this potential confound. It is also possible that the decision to include only a single talker in the lexical competition task in Experiments 3, 6, and 7 rendered talker information less influential during spoken word recognition (Pilotti et al., 2000).

5.4 Combined Analyses

In order to investigate in more detail how TSEs in recognition memory were altered by the presence of within- and between-talker variability during study, as well as determining whether talker variability during study affected lexical competition effects, data from Experiments 2, 3, 6, and 7 were combined and analysed with variability (none, within-talker, or between-talker) included as a between-participant or within-items factor. In Experiments 2 and 3 participants heard only a single token of each novel word spoken by a single talker (*no variability*) during the study-phase of the experiment. In Experiment 6 participants heard 18 different tokens from a single talker during study (*within-talker variability*), and in Experiment 7 participants heard 9 tokens from two talkers (1 male and 1 female) for each novel word (*between-talker variability*). It is important to note that all experiments used the same items and counterbalancing for the test-phase of the experiments and that participants completed tests of talker-specificity and lexical

competition effects at the same time points in each of the four experiments. It was necessary to include both Experiments 2 and 3 in the combined analysis since, as can be seen in Table 5.2, Experiment 2 did not include confidence rating or a male/female categorisation task, and so could not be compared to Experiments 6 and 7 on these measures, whilst Experiment 3 used pause detection rather than lexical decision, and thus cannot be compared with the lexical decision data from Experiments 6 and 7. Thus, Experiments 3, 6, and 7 were included in the combined analysis of TSEs in recognition memory whilst Experiments 2, 6, and 7 were included in the combined analysis of lexical competition effects, resulting in a comparison of no, within-talker, and between-talker variability for both sets of data.

Table 5.2. Summary of tasks completed in each of the experiments included in the cross-experiment analyses

Exp	Variability at study	Test of lexical competition	Old/new categorization	Confidence ratings	Male/female categorization
2	None	Lexical decision	✓	×	×
3	None	Pause detection	✓	✓	✓
6	Within-talker	Lexical decision	✓	✓	✓
7	Between-talker	Lexical decision	✓	✓	×

5.4.1 Talker-specificity effects

All of the TSE data were analysed in repeated-measures ANOVAs with variability (none, within-talker, *vs.* between-talker), day (1, 2, *vs.* 8), and test-phase talker (same *vs.* different) included as within-participant variables, and list (1 *vs.* 2) included as a between-participants variable in order to reduce the estimate of random variation (Pollatsek & Well, 1995). Only main effects and interactions involving the factor *variability* are reported below as these are the effects of primary interest in these analyses.

For the d′ data (Figure 5.7a) from the *old/new categorisation* task there was no main effect of variability, $F < 1$. However, there was a significant interaction between variability and test-phase talker, $F(2,103) = 7.66$, $p = .001$, $\eta_p^2 = .13$. This interaction remained significant when no-variability was compared to both within-talker variability, $F(1,75) = 8.80$, $p < .01$, $\eta_p^2 = .12$, and between-talker variability, $F(1,74) = 10.99$, $p = .001$, $\eta_p^2 = .13$, but was not significant in the comparison of

within- and between-talker variability, $F < 1$, suggesting that introducing any type of variability during study resulted in smaller TSEs in the old/new categorization task. The interaction between variability and day was not significant, $F(4,206) = 1.11$, *ns*, nor was the three-way interaction between variability, test-phase talker and day, $F < 1$, indicating that the interactions observed between variability and test-phase talker were stable across the course of a week.

In the comparison of no variability and within-talker variability further analysis revealed that the main effects of variability was significant only for different-talker items, $F(1,75) = 4.43$, $p < .05$, $\eta_p^2 = .06$, suggesting that including within-talker variability during study increased correct classification of different-talker items in the old/new categorization task (see Table 5.3). The main effect of variability was not significant for same-talker items, $F(1,75) = 1.50$, *ns*. Thus, it seems that including within-taker variability during the study phase of the experiment resulted in representations of the novel words that were more robust to changes in talker at test.

Conversely, the comparison of no variability and between-talker variability revealed a main effect of variability only for same-talker items, $F(1,74) = 8.93$, $p < .01$, $\eta_p^2 = .11$, not for different-talker items, $F(1,74) = 1.18$, *ns*, with recognition accuracy for same-talker items being lower in the between-talker variability experiment relative to the no-variability experiment (Table 5.3) suggesting that talker-information was less influential in the recognition process. However it is important to note that participants heard only 9 tokens spoken by each individual talker in Experiment 7 whereas participants in Experiment 3 were exposed to 18 tokens spoken by a single talker. Thus, poorer categorisation of same-talker items in Experiment 7 is perhaps unsurprising given the differences in exposure during the study phase of the experiment. However if this was the case then we may also have expected a main effect of variability for different-talker items. In other words, assuming that fewer exposures to each individual talker during study results in less robust representation of talker information associated with each novel item then it might be expected that a change in talker at test should have had a smaller effect on old/new categorisation responses and resulted in higher d′ values for different-talker items in Experiment 7 compared to Experiment 3. This was not the case. Rather, it seems more likely that the decrease in accuracy for same-talker items in the

between-talker variability experiment results from the introduction of multiple talkers during study.

For β data (Figure 5.7b) comparison of Experiments 3, 6, and 7 did not reveal a significant main effect of variability, $F(2,103) = 1.84$, *ns*, nor did variability interact with day, $F < 1$. There was however a significant interaction between variability and test-phase talker, $F(2,103) = 6.20$, $p < .01$, $\eta_p^2 = .11$. This interaction remained significant in the comparisons of no variability and between-talker variability, $F(1,74) = 5.41$, $< .05$, $\eta_p^2 = .07$, and within- and between-talker variability, $F(1,57) = 12.41$, $p = .001$, $\eta_p^2 = .18$. The interaction between variability and test-phase talker was not significant in the comparison of no- and within-talker variability, $F(1,75) = 2.23$, *ns*. These findings suggest that the size of the TSEs for β values decreased significantly only when multiple talkers were introduced during the study-phase of the experiment (Figure 5.7b). As in the d′ analysis the three-way interaction between variability, test-phase talker, and day was not significant, $F < 1$, suggesting that the interactions observed between variability and test-phase talker were stable across the three test sessions.

In the comparison of no variability and between-talker variability further analysis indicated that the main effect of variability was significant for different-talker items, $F(1,74) = 4.51$, $p < .05$, $\eta_p^2 = 06$, but not for same-talker items, $F < 1$. Note this is the opposite pattern of data to that observed for the d′ data when comparing these two experiments. Nonetheless, further analysis comparing data from the within and between-talker variability experiments also revealed a significant main effect of variability for different talker items, $F(1,57) = 10.10$, $p < .01$, $\eta_p^2 = .15$, but not same-talker items, $F < 1$. In both cases the main effect of variability for different-talker items reflects lower β values in the between-talker experiment, suggesting that after training in which each novel item was heard in the two voices participants are less biased to classify different-talker items as new. It is interesting that hearing two talkers per item during study does not also appear to decrease the bias in classifying same-talker items as new.

Analysis of AUC values (Figure 5.7c) once again revealed a non-significant main effect of variability, $F < 1$. As in d′ analyses the only significant interaction was between variability and test-phase talker, $F(2,101) = 8.97$, $p < .001$, $\eta_p^2 = .09$, with this interaction remaining significant when comparing no variability to either

within-talker variability, $F(1,74) = 13.03$, $p = .001$, $\eta_p^2 =. 15$, or between-talker variability, $F(1,74) = 13.42$, $p < .001$, $\eta_p^2 = .16$, but not when comparing within- and between-talker variability, $F < 1$. These findings add support to conclusions draw from the d′ data, indicating that introducing any time of variability during the study-phase resulted in smaller TSEs for those items in later test session.

Further analysis of the no variability and between-talker variability experiments revealed a significant main effect of variability only for same-talker items, $F(1,72) = 7.43$, $p < .01$, $\eta_p^2 = .09$, not for different-talker items, $F(1,72) = 2.62$, *ns*. This is consistent with the d′ data, indicating that introducing multiple talkers for each item during study not only decreases the accuracy with which same-talker items are classified as old or new, it also, unsurprisingly, decreases participants confidence in these decisions.

Comparison of no variability and within-talker variability also revealed a significant main effect of variability for same-talker items, $F(1,74) = 4.97$, $p < .01$, $\eta_p^2 = .06$, but not different-talker items, $F(1,74) = 2.88$, $p = .094$, $\eta_p^2 = .04$. Note that this is the opposite to d' data which showed a main effect of variability only for different-talker items. In this case, it appears that whilst the introduction of within-talker variability during study does not affect accuracy in categorising same-talker items as old or new compared to the no-variability experiment, it does decrease participants' confidence in making these decisions.

Finally, variability did not produce a significant main effect in the male/female categorisation task, $F < 1$, nor were any of the interaction involving variability significant in this analysis. Note, only Experiments 3 and 6 were included in this comparison since participants heard each novel word in both a male and a female voice during the study phase of Experiment 7 (between-talker variability), and thus a source memory judgment about the study talker could not be made.

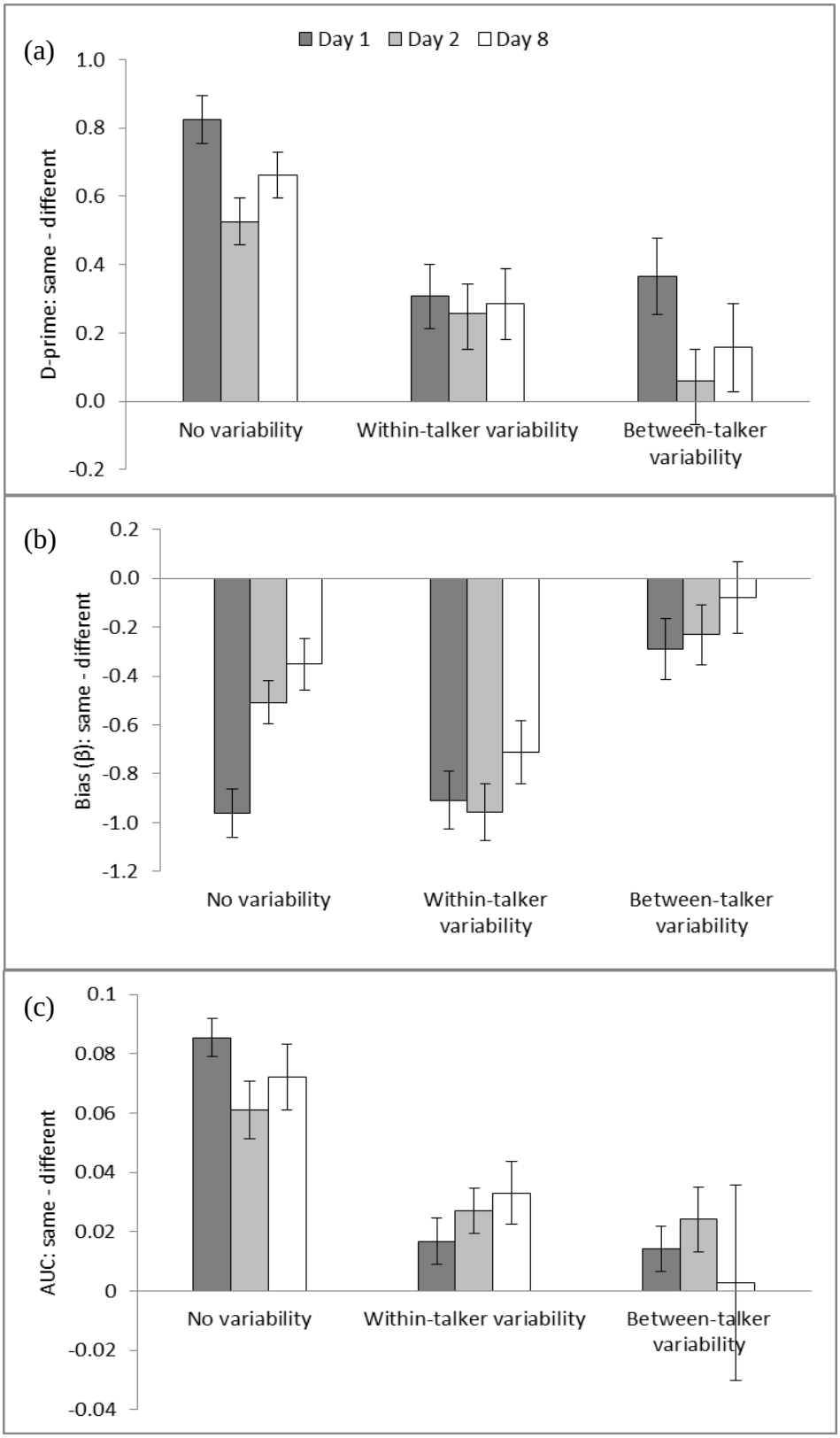

Figure 5.7. Same-talker minus different-talker scores in each test session for (a) d-prime, (b) beta, and (c) AUC scores in Experiments 3 (no variability), 6 (within-talker variability), and 7 (between-talker variability). Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

Table 5.3. Mean scores for all tasks examining TSEs, split by test session, and by test-phase talker. Standard error of the mean is provided in parentheses. Experiment 3 = no variability; Experiment 6 = within-talker variability; Experiment 7 = between-talker variability

Measure	**Exp**	**Day 1**			**Day 2**			**Day 8**		
		Same	*Different*	*Same-Different*	*Same*	*Different*	*Same-Different*	*Same*	*Different*	*Same-Different*
D-prime (old/new categorisation)	3	2.55 (.07)	1.73 (.07)	.82	2.30 (.08)	1.78 (.06)	.52	2.41 (.07)	1.75 (.07)	.66
	6	2.36 (.09)	2.05 (.10)	.31	2.31 (.09)	2.06 (.09)	.25	2.19 (.11)	1.90 (.10)	.29
	7	2.32 (.12)	1.95 (.11)	.37	2.08 (.09)	2.02 (.10)	.06	1.96 (.13)	1.80 (.13)	.16
Beta (old/new categorisation)	3	.84 (.08)	1.79 (.12)	-.95	.95 (.08)	1.46 (.10)	-.51	1.08 (.10)	1.43 (.11)	-.35
	6	.87 (.09)	1.78 (.15)	-.91	.85 (.08)	1.80 (.16)	-.95	1.12 (.11)	1.84 (.14)	-.72
	7	.88 (.13)	1.17 (.12)	-.29	1.00 (.11)	1.23 (.13)	-.23	1.20 (.15)	1.28 (.14)	-.08
AUC (confidence ratings)	3	.97 (.01)	.89 (.01)	.08	.95 (.01)	.88 (.01)	.07	.95 (.01)	.88 (.01)	.07
	6	.94 (.01)	.92 (.01)	.02	.94 (.01)	.91 (.01)	.03	.92 (.01)	.89 (.01)	.03
	7	.93 (.03)	.92 (.03)	.01	.93 (.02)	.91 (.03)	.02	.90 (.04)	.90 (.04)	.00
Error (% male/female categorisation)	3	16.3 (2.4)	42.9 (3.5)	-26.6	18.8 (2.1)	47.9 (2.6)	-29.1	22.2 (2.4)	51.4 (3.0)	-29.2
	6	15.5 (2.4)	42.8 (3.4)	-27.3	16.6 (2.8)	48.7 (3.0)	-32.1	23.7 (3.2)	46.0 (3.8)	-22.3

5.4.2 Lexical competition effects

As well as investigating changes in lexical competition data as a result of differences in talker variability during the study-phase of the experiments, an additional motivation for conducting a combined analysis of the lexical competition data was the intriguing (but mostly non-significant) patterns of data observed in the lexical decision tasks from Experiment 2, 6, and 7 suggesting that talker information may influence the amount of competition observed between novel nonwords and their phonologically-similar basewords at delayed test points. As noted above in the discussion of Experiment 7 one limitation of the tests of lexical competition reported in this thesis is that dividing the items into same-talker versus different-talker test basewords resulted in measures of same-talker and different-talker lexical competition effects that relied on analysis of RTs to only 12 items. It seems reasonable to suggest that having such a small number of items will have limited the power of the analysis. To overcome this problem data from Experiments 2, 6, and 7 were combined to increase the power in the analyses through the increase in sample size.

Data from the 48 basewords were analysed in a repeated-measures ANOVA with factors variability (none, within-talker, *vs.* between-talker), baseword type (test *vs.* control) and day (1, 2, *vs.* 8). With regards to our first question (Does talker variability during study affect the pattern of lexical competition effects for newly learned words?) this analysis showed that the main effect of variability was non-significant by participants $F_1(2,83) = 2.43$, *ns*, but significant by items, $F_2(2,84) = 97.85$, $p < .001$, $\eta_p^2 = .70$. The interaction between variability and day was also significant only in the by-items analysis, $F_2(4,168) = 3.30$, $p < .05$, $\eta_p^2 = .07$. These differences between by-participants and by-items analyses may stem from the fact that variability is a within-items variable, as compared to a between-participants variable, and so the by-items analysis may be more powerful in detecting effects of variability. No other interactions involving variability approached significance. Given the differences between by-participants and by-items analyses it is difficult to draw any firm conclusions with regards to the effects of variability on lexical competition measures.

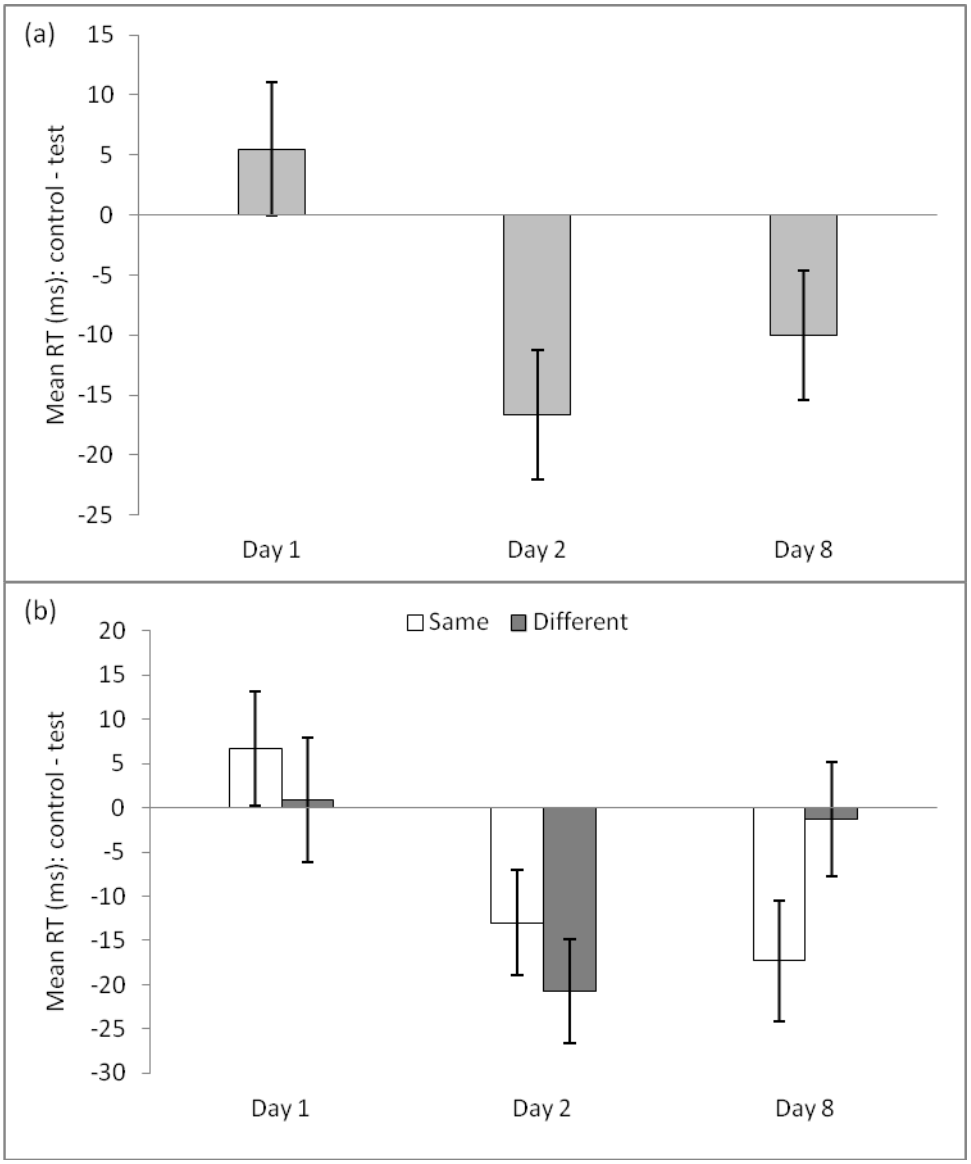

Figure 5.8. (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) basewords in the lexical decision task. (b) Lexical decision data split according to whether the test baseword was spoken in the same voice that the corresponding novel word was trained in, or a different voice. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

This initial analysis also revealed, unsurprisingly, a main effect of day, $F_1(2,66) = 20.77$, $p < .001$, $\eta_p^2 = .20$, $F_2(1.7,70.1) = 74.59$, $\eta_p^2 = .64$, indicating that RTs decreased across test sessions. The main effect of baseword type was also marginally significant by participants, $F_1(1,83) = 3.80$, $p = .055$, $\eta_p^2 = .04$, and was significant by items, $F_2(1,42) = 6.44$, $p < .05$, $\eta_p^2 = .13$. More importantly, the critical interaction between day and baseword type was significant, $F_1(2,166) = 5.11$,

$p < .01$, $\eta_p^2 = .06$, $F_2(2,84) = 3.18$, $p < .05$, $\eta_p^2 = .07$. Further analysis revealed that there was no significant difference between RTs to test and control basewords on Day 1, $F < 1$, but that this difference was significant on Day 2, $F_1(1,85) = 9.67$, $p < .01$, $\eta_p^2 = .10$, $F_2(1,42) = 11.23$, $p < .01$, $\eta_p^2 = .21$, and marginally significant on Day 8, $F_1(1,84) = 3.79$, $p = .055$, $\eta_p^2 = .04$, $F_2(1,42) = 3.64$, $p = .063$, $\eta_p^2 = .08$, suggesting that lexical competition was absent immediately after study, emerged on Day 2, and was retained (to some degree) over the course of a week (Figure 5.7b), consistent with previous studies examining word learning in adults (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008). Notably, in this analysis there was no significant facilitation effect for test basewords on Day 1 (*i.e.*, faster RTs to test than control basewords), arguing against the suggestion that lexical competition effects may be absent immediately after study as a result of phonological priming between the studied novel nonwords and the phonologically-similar basewords heard in the test of lexical competition. If this were the case then significant priming effects should have been observed on Day 1.

In order to address the second question (whether a combined analysis would clarify the tentative pattern of talker-specific lexical competition effects observed in individual experiments) the lexical competition data were broken down further, with the test basewords being divided into sets of same- and different-talker items depending on whether they were heard in the same voice that the studied novel nonword had been encountered in during the phoneme monitoring task or not (Figure 5.7b). Data were analysed in a repeated-measures ANOVA with factors variability (none, within-talker, *vs.* between-talker), baseword-type (same-talker, different-talker, *vs.* control), and day (1, 2 *vs.* 8).

This analysis revealed a significant main effect of day, $F_1(2, 164) = 19.62$, $p < .001$, $\eta_p^2 = .19$, $F_2(2,82) = 48.86$, $p < .001$, $\eta_p^2 = .54$, but a non-significant main effect of baseword-type, $F_1(1.7,145.3) = 1.24$, *ns*, $F_2(1.6, 3.6) = 2.27$, *ns*. Nevertheless, the interaction between day and baseword-type was significant, $F_1(3.2,266.4) = 2.63$, $p < .05$, $\eta_p^2 = .03$, $F_2(3.1,130.7) = 2.17$, $p < .05$, $\eta_p^2 = .07$. In order to determine whether the lexical competition effects that were observed in the main analysis on Days 2 and 8 were significant for same-talker and/or different-talker items on each day further analyses were carried out. On Day 2 there was a

significant main effect of baseword-type, $F_1(2,168) = 4.48$, $p < .05$, $\eta_p^2 = .05$, $F_2(1,84) = 2.76$, $p = .069$, $\eta_p^2 = .06$. Further analyses revealed differences between RTs to same-talker items and control items, $F_1(1,84) = 3.99$, $p < .05$, $\eta_p^2 = .05$, $F_2(1,42) = 4.01$, $p = .052$, $\eta_p^2 = .09$, as well as between different-talker items and control items, $F_1(1,84) = 8.88$, $p < .01$, $\eta_p^2 = .10$, $F_2(1,42) = 8.37$, $p < .01$, $\eta_p^2 = .17$, suggesting that lexical competition was observed for all items on Day 2, regardless of whether the baseword was spoken by the same talker that the corresponding novel nonword had been studied in. In support of this suggestion there was no difference between RTs to same- and different-talker items in the Day 2 test session, $F_1(1,84) = 1.02$, *ns*, $F_2 < 1$. On Day 8 the main effect of baseword-type was marginally significant by-participants, $F_1(2,166) = 2.40$, $p = .094$, $\eta_p^2 = .03$, but was significant by-items $F_2(2,82) = 6.61$, $p < .001$, $\eta_p^2 = .14$. Further analysis revealed that at this time point only RTs to same-talker items differed significantly from control items, $F_1(1,83) = 6.00$, $p < .05$, $\eta_p^2 = .05$, $F_2(1,41) = 9.31$, $p < .01$, $\eta_p^2 = .19$. The difference between RTs to different-talker and control items was no longer significant, $F < 1$. Moreover, the difference in RTs for same-talker and different-talker items approached significance by-participants, $F_1(1,83) = 2.22$, $p = .14$, $\eta_p^2 = .03$, and was significant by items, $F_2(1,42) = 9.06$, $p < .01$, $\eta_p^2 = .18$, at this time point. Together these findings suggest that one week after learning a set of novel nonwords, information about the voice in which these items had been studied affected the degree to which they engaged in competition with similar sounding existing words.

5.5 Chapter Summary and Discussion

The aim of the two experiments reported in this chapter was to explore the effects of talker variability on the time-course of TSEs and lexical competition effects for newly-learned words. The combined analyses indicated that there were no main effects of variability. In other words, the amount of talker-variability during study did not affect the overall level of performance in any task. Variability did however alter the size of the TSEs observed in d′, β, and AUC data (see Table 5.4. for a summary of the combined analyses).

In the d′ and AUC data TSEs were smaller following any kind of variability. In the AUC data these smaller TSEs resulted from lower confidence ratings for same-talker items following both within- and between-talker variability. Likewise,

following between-talker variability d′ scores were lower for same-talker items, presumably because talker information was less useful during recognition of words that were originally studied in more than one voice. In contrast, following within-talker variability there appeared to be an increase in d′ scores for different-talker items but no change in performance for same-talker items. This latter finding suggests that the introduction of within-talker variability during study resulted in recognition performance being more robust to changes in talker, a finding that was mirrored in the β data, in which participant showing less of a bias to say that different-talker items were new following within-talker variability during study. Likewise, β data also suggested that participants showed less of a bias to label different talker items as new following between-talker variability during study. Thus, the effects of within- and between-talker variability are rather complicated, with different experimental measures indicating different effects of introducing different types of variability during study.

Table 5.4. Summary of findings from the combined analyses for d′, β, AUC, and male/female categorization data comparing no-variability, within-talker variability, and between-talker variability (Experiments 3, 6, and 7 respectively). Each cell indicates whether a difference arose for same-talker or different-talker items, and which condition had a higher mean.

Task	None *vs.* Within	None *vs.* Between	Within *vs.* Between
Old/new categorization (d′)	Different-talker None < Within	Same-talker None > Between	No difference
Old/new categorization (β)	No difference	Different-talker None > Between	Different-talker Within > Between
Confidence ratings (AUC)	Same-talker None > Within	Same-talker None > Between	No difference
Male/female categorisation	No difference	NA	NA

Nonetheless, whilst variability during study appeared to alter the size of TSEs during recognition of newly-learned words, it did not appear to alter the *time-course* of TSEs. Evidence supporting this claim comes from the non-significant interactions involving day in the combined analyses. Thus, whilst the data suggest that talker information plays a smaller role in recognition memory judgments following exposure to more variable study tokens, these effects appeared to be stable over time.

Notably variability did not affect the size of the same-talker advantage in the male/female categorization task, indicating that talker information was successfully encoded and stored despite the presence, or absence, of variability in the input, and

that when required this information was equally accessible in Experiments 3 and 6. It is important to note however that the male/female categorization task could not be used in Experiment 7 since all items were studied in both a male and a female voice. Thus, it remains possible that explicit access to information about the study talker may be altered once each item is heard in multiple voices during study.

In contrast to the stability of TSEs over time, lexical competition effects in the combined analysis did not emerge until Day 2, with the critical interaction between day and word-type being statistically significant in this more powerful analysis. These findings are consistent with previous research examining word learning in both adults (*e.g.*, Dumay & Gaskell, 2007) and children (Henderson et al., submitted-a, submitted-b). Lexical competition effects also remained close to significance at the one week retest, again consistent with previous research (*e.g.*, Tamminen & Gaskell, 2008). As noted in Chapter 3 the different time-courses of TSEs and lexical competition effects suggest that recognition memory and lexical competition rely on different representational systems. However, the combined analysis also revealed that the lexical competition effects appeared to be talker-specific on Day 8, one week after the novel words were initially learned. More specifically, lexical competition effects were observed only when the existing word (*e.g.*, *biscuit*) was heard in the same voice that the corresponding novel nonword (*e.g.*, *biscal*) had been studied in. This finding was unexpected given the assumption that lexical competition relies on overlap between abstract phonological representations. However the replication of a trend towards only same-talker lexical competition on Day 8 across multiple experiments (2, 3, and 7), as well as in the combined analysis, suggests that this finding requires further discussion.

One possible explanation for the talker-specific lexical competition effects on Day 8 is that exposure to the existing basewords in the lexical decision task on Days 1 and 2 may result in the formation of episodic traces of these items, and that it is these representations in the episodic subsystem that drive the talker-specific lexical competition effects in conjunction with the episodic traces of the newly-learned words. However, if this explanation were correct then talker-specific lexical competition should have been observed on Day 2 as well as Day 8 since episodic traces of the existing words should have been established during the lexical decision task on Day 1. Moreover, given that lexical competition requires interactions between similar-sounding words during spoken word recognition it seems unlikely that episodic traces, presumably

maintained in the hippocampal system, which are sparsely coded and pattern-separated (Bakker, et al., 2008) will be capable of supporting this type of interaction. It is more likely that distributed, overlapping representations are required (as in the neocortical system of the CLS framework, McClelland et al., 1995). Thus it seems unlikely that talker-specific lexical competition effects would be able to arise from the interaction of episodic representations of existing and novel words in the episodic subsystem.

An alternative suggestion is that talker-specific lexical competition may arise due to co-activation of episodic and abstract representations of the novel nonwords, which would presumably result in stronger activation of same-talker items compared to different-talker items, making novel nonwords stronger competitors for their phonologically-similar basewords on same-talker trials. Evidence suggesting that the rapidly-formed episodic representations are maintained for at least a week after initially studying the novel nonwords comes from the finding that TSEs in the old/new categorisation task remained significant or marginally significant at all time points in Experiments 1-6. On the other hand, data from the confidence ratings suggest that recollection processes, which are assumed to rely heavily on activation in the hippocampus (Davachi, Mitchell, & Wagner, 2003; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Ranganath et al., 2004), decreased over the course of a week in Experiments 3, 6, and 7. Likewise, the accuracy with which participants were able to explicitly access information about the study talker of each item decreased over time in Experiments 3 and 6. These latter two findings suggest that access to source information about the study phase of the experiment decreased over time, possibly as a result of trace decay in the episodic subsystem. Thus, if episodic traces decay over time but talker-specific lexical competition effects arise due to co-activation of episodic and abstract representations of the novel nonwords, then stronger same-talker lexical competition effects should have been observed on Day 2 compared to Day 8. This was not the case.

However, it may be argued that whilst an initial consolidation period is required in order to establish and stabilise an abstract phonological representation of a new word so that it is robust enough to engage in lexical competition with similar-sounding existing words (accounting for the emergence of lexical competition effects on Day 2, but not on Day 1), consolidation processes may continue to further strengthen and enhance the newly formed abstract representations over the course of a week. Moreover, repetition of the novel words during the Day 2 re-test may trigger

reconsolidation processes (Lee, 2009; Nader & Einarsson, 2010; Stickgold & Walker, 2007), further strengthening abstract phonological representation of the novel words. If this is the case then novel words may be activated more 'automatically' during the one week re-test, possibly allowing extra time, and perhaps extra processing resources, for talker information in the episodic representations to be integrated with the retrieved abstract representation of a novel word, allowing same-talker novel nonwords to act as stronger competitors during recognition of the existing baseword only at the most delayed test point. Thus, it remains possible that talker-specific lexical competition effects could have arisen due to a combination of activation of representations of the novel nonwords in both the episodic and abstract subsystems.

An alternative explanation may be that talker information, as well as phonological information, is consolidated and strengthened over time, resulting in talker-specific lexical competition effects at delayed test points. However, if memory for talker information was strengthened over time then larger TSEs should also have been observed in the old/new categorisation task on Day 8. This was clearly not the case. However, it may be that whilst consolidation processes strengthen the abstract phonological representation of a new word, resulting in improved free and cued recall of newly learned words (Brown, et al., in press; Dumay & Gaskell, 2007; Henderson, et al., submitted-a, submitted-b; Tamminen, et al., 2010) as well as faster recognition (Snoeren, et al., 2009) and repetition of newly learned words (Davis & Gaskell, 2009; Gagnepain, Henson, & Davis, 2010), the consolidation processes involved in stabilising information about the study talker of an item may instead be important in generating robust links between talker information and specific phonological representations rather than strengthening memory for the talker information *per se*. Moreover, the time-course with which these two types of information are consolidated may differ. Phonological representations must be stabilised first during initial periods of sleep-associated offline consolidation in order for lexical competition effects to be observed on Day 2. It is only after these phonological representations have been established and stabilised that further consolidation processes may then link these representations to specific talker information, thus accounting for talker-specific lexical competition effects only at the most delayed test point in our experiments.

Evidence suggesting that abstract and episodic information are processed differently comes from work by Marsolek and colleagues (Burgund & Marsolek, 1997; Gonzalez, Cerveza-Crespo, & McLennan, 2010; Gonzalez & McLennan, 2007, 2009; Marsolek, 1999, 2004; Marsolek & Burgund, 2008; Marsolek, Kosslyn, & Squire, 1992; Marsolek, Squire, Kosslyn, & Lulenski, 1994) demonstrating hemispheric differences in processing of these two types of information. The LH is assumed to specialise in processing of abstract information (in this case the phonological information), and the RH is assumed to specialise in processing of episodic information (the talker-specific information). Moreover, evidence from patients with phonagnosia who show impairments in voice recognition but intact comprehension of phonological information (Vanlancker, Cummings, Kreiman, & Dobkin, 1988) suggests that phonological and talker information may be stored separately in lexical memory. Thus, it does not seem implausible to suggest that consolidation processes may affect these two types of information differently.

Nonetheless, what is particularly striking about the pattern of lexical competition data in the combined analysis is that RTs to same- and different-talker items were equivalent on Day 2, with lexical competition effects emerging for both types of item, whereas by Day 8 only same-talker items showed significant lexical competition effects. Whilst the two explanations outlined above are capable of accounting for greater lexical competition effects for same- compared to different-talker items, neither fully accounts for the *absence* of lexical competition effects for different-talker items on Day 8. In order to account for this pattern of data it may be necessary to assume that once talker information is able to influence lexical competition effects a mismatch in talker information between the existing and novel words (*e.g.*, *biscuit* and *biscal*) inhibits competition between these items. This assumption could be applied to both of the explanations outlined above.

CHAPTER 6: GENERAL DISCUSSION

The main aims of this thesis were to examine the degree to which episodic and abstract representations are involved in spoken word recognition during word learning and to investigate whether the contributions of these two types of representation change over time, possibly as a result of consolidation processes. Specifically, the experiments reported in this thesis examined whether information about the study talker of a novel word affected later recognition of that item as well as the degree to which the word engaged in lexical competition with phonologically-similar existing words. By examining the time-course of TSEs for these two processes concurrently using the same set of stimuli it was hoped that it would be possible to determine whether consolidation processes strengthen memory for all types of information, or whether consolidation is a more selective process that is involved in the generation of more robust abstract phonological representations.

6.1 Summary of findings

6.1.1 Talker-specificity effects in recognition memory

All experiments reported in this thesis demonstrated that phonological representations of novel words were rapidly established, and were able to support accurate recognition of the newly-learned items immediately after study, consistent with a number of previous studies examining word learning in both adults (Davis et al., 2009; Dumay & Gaskell, 2007; Dumay et al., 2004; Gaskell & Dumay, 2003; Leach & Samuel, 2007; Snoeren et al., 2009; Tamminen & Gaskell, 2008; Tamminen et al., 2010) and children (Brown et al., in press; Carey, 1978; Frazier Norbury et al., 2010; Henderson et al., submitted-a, submitted-b; Markson & Bloom, 1997). Moreover, talker-specific information was encoded and stored during study and was used to aid later recognition of the newly-learned words, as evidenced by higher recognition accuracy for same-talker items in the old/new categorisation task on Day 1 of all experiments. The presence of TSEs in recognition memory in the immediate test session on Day 1 is consistent with prior research showing that talker information can affect recognition and processing of existing and novel words when participants are tested immediately after study (Bradlow et al., 1999; Craik & Kirsner, 1974; Creel et al., 2008; Creel & Tumlin, 2009, 2011; Goh, 2005; Goldinger, 1996; Goldinger et al., 1999; McLennan & Luce, 2005; Schacter &

Church, 1992; Sheffert, 1998), as well as research suggesting that other surface-form variables affect recognition of recently studied items (*e.g.*, font-specific memory; see Tenpenny, 1995, for a review).

Experiment 2 built on these previous findings by showing that talker-specific information continued to affect recognition of newly-learned words up to one week after study, with no significant decrease in the contribution of talker information to recognition memory. This finding suggests that episodic representations of newly-learned words can be maintained in memory for at least a week after initially learning the items, at least under conditions where there is little variability in the input. Experiment 3 replicated this finding, and also examined explicit memory for information about the study talker. Participants showed a decrease in their ability to explicitly access information about the study talker as the week progressed. This finding suggests that information about the study talker is gradually lost as a result of trace decay of episodic representations over time. In addition, analysis of confidence ratings indicated that as the week progressed participants became gradually less reliant on recollection processes and more reliant on familiarity processes when making their old/new categorization decisions. This decrease in reliance on recollection processes over time also suggests that highly-detailed episodic representations become less involved in recognition of newly learned words over time. Given evidence suggesting that recollection is highly dependent on hippocampal regions (Davachi et al., 2003; Eldridge et al., 2000; Ranganath et al., 2004), and that familiarity mechanisms are more dependent on surrounding MTL regions (Brandt et al., 2009; Elfman et al., 2009; Henson et al, 2003; Ranganath et al., 2004) this finding appears to be consistent with the predictions of a hybrid model of lexical representation in which the contribution of episodic representations decreases over time as abstract representations of newly-learned words are strengthened and stabilised (assuming that MTL regions surrounding the hippocampus support abstract representation, as suggested by a CLS model; McClelland et al., 1995).

The maintenance of TSEs for novel words over the course of a week in the old/new categorization task in Experiments 2 and 3 contrasts with the decrease in TSEs observed by Goldinger (1996) for existing words in a similar old/new categorization task. Experiments 4 and 5 directly compared the time-course of TSEs in recognition memory for existing and novel words to determine whether the

differences between Experiments 2-3 and Goldinger's study were simply due to methodological differences, or whether the pattern of TSEs was altered by the presence or absence of pre-established representations of a word in memory prior to the experiment. Data revealed that TSEs were equivalent in size for existing and novel words immediately after study and that the time-course of TSEs in recognition memory was not significantly different for the two types of item. However, further exploration of the data suggested that there were some subtle differences between recognition of existing and novel words in the one-week re-test. Existing words, unlike novel words, did not exhibit a robust same-talker advantage in the delayed test session. Evidence that source recollection of information about the study talker decreased at the same rate for existing and novel words indicates that episodic traces decayed at the same rate regardless of whether the item has been encountered prior to the experiment or not. Thus, the absence of TSEs for existing but not novel words in recognition memory on Day 8 must result from a greater contribution from pre-established, presumably abstract, representations during recognition of existing words at the delayed test points. However, as stated in Chapter 4, this finding must be interpreted with caution due to the lack of significant interactions in the main analysis.

Finally, Experiments 6 and 7 examined whether variability in the input during study was an important factor in determining how long talker information was retained in memory and used to aid recognition of newly-learned words. According to exemplar models of lexical representation the greater the number of different episodic traces in memory for each item the less talker-specific the retrieved representation should be. Consistent with this prediction Experiments 6 and 7, which examined the effects of within- and between-talker variability respectively, revealed significantly smaller TSEs than Experiment 3 in which there was no variability in the study tokens. These reduced TSEs did not change significantly in size across the course of a week in either experiment, suggesting once again that highly-detailed episodic representations are retained in memory for at least a week after study. However, further analysis revealed non-significant TSEs in both the Day 2 and Day 8 delayed test sessions following between-talker variability during study. This pattern of data is remarkably similar to that observed in Experiment 5 for existing words, and is potentially indicative of a role of stimulus variability in the generation of robust abstract representations of newly-learned words, although clearly this

suggestion requires further investigation due to the lack of a significant interaction between day and test-phase talker in the main analysis.

6.1.2 Lexical competition effects

In comparison to TSEs, lexical competition effects were absent on Day 1, immediately after the novel nonwords were studied, but emerged one day later after a period of sleep-associated offline consolidation, consistent with a number of previous studies examining the emergence of lexical competition for novel words in both adults (Davis, et al., 2009; Dumay & Gaskell, 2007; Dumay, et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008) and children (Henderson, et al., submitted-a, submitted-b). The finding that lexical competition effects emerged at a different time-point to TSEs in recognition memory is critical as it suggests that different processing mechanisms, and potentially different types of representation, underlie these two different effects. This point will be discussed further in section *6.1.3* below.

In contrast to the highly-detailed episodic representations that are assumed to underlie TSEs, the emergence of lexical competition is assumed, according to most models of spoken word recognition such as the cohort model (Marslen-Wilson & Zwitserlood, 1989), or TRACE (McClelland & Elman, 1986), to rely on interactions between abstract representations of phonologically similar words. Evidence supporting the suggestion that lexical competition effects involve activation of abstract representations comes from the finding that significant lexical competition effects were observed for both same- and different-talker items on Day 2 of the combined analysis. However, the combined analysis also suggested that lexical competition effects were talker-specific in the one-week retest. As discussed at the end of Chapter 5 there are two possible explanations of these findings. Either talker-specific lexical competition effects arise due to co-activation of abstract and episodic representations of the novel nonwords, allowing same-talker novel nonwords to become stronger competitors than different-talker novel nonwords, or talker information (as well as phonological information) may be consolidated in memory over time.

Interestingly, a previous study by Creel et al. (2008) suggests that that talker-specific lexical competition effects can be observed for both existing and novel words. Thus, it does not appear that these effects are limited to newly-learned items

that presumably rely on episodic representations to a greater extent than existing words since existing words already have pre-established, presumably abstract, phonological representations in lexical memory. However, in both Creel et al.'s study and the experiments reported in this thesis the stimuli were repeated a number of times in a single voice during the study phase of each experiment. Even in Experiment 7 participants heard each novel nonword spoken nine times by each talker during study. This repetition may have enhanced information about the talker associated with each word and resulted in greater use of this information during spoken word recognition than would typically be the case when processing everyday speech. Further research examining talker-specific lexical competition effects is needed in order to establish whether similar effects would be observed in cases where each word was heard only once prior to the test phase of the experiment.

6.1.3 Main advances in this thesis

The key finding in this thesis is that TSEs and lexical competition effects for newly learned words follow different time-courses. TSEs are present immediately after novel words have been studied and, depending on the amount of variability in the input, can remain stable for up to one week post-study. Lexical competition effects on the other hand are absent immediately after new words have been learned, but emerge one day later follwoing a period of sleep-associated offline consolidation. Evidence suggesting that the time-course of TSEs and lexical competition effects are independent comes from the lack of correlations between TSEs and lexical competition effects in almost all experiments. Moreover, the finding that TSEs in recognition memory did not decrease as lexical competition effects increased between Days 1 and 2 (at least when there was limited variability in the study tokens) suggests that the processes whereby representations underlying lexical competition effects are strengthened and stabilised in memory over time do not influence the rate of decay of episodic traces.

Further evidence supporting the presence of different systems underlying TSEs and lexical competition effects comes from the finding that variability during study and sleep-associated offline consolidation appeared to affect these two measures differently. Variability influenced the size of TSEs in recognition memory (*i.e.*, the more variability included during study the less robust the TSEs in recognition memory), but did not appear to impact upon the pattern of lexical competition effects

(Figures 3.2b and 5.6b show almost identical patterns of data). In comparison, sleep-associated offline consolidation appeared to be important for the emergence of lexical competition effects one day after the new words were studied, but did not appear to have robust effects on TSEs in recognition memory since these effects were, for the most part, stable over time.

Taken together, the different time-courses of TSEs and lexical competition effects suggest that these two effects may be dependent on different processing mechanisms and potentially different types of representations. More importantly, the independence of the time-courses of TSEs and lexical competition effects suggests that the two systems underlying these effects must be able to co-exist. This latter suggestion is consistent with hybrid models of lexical representation which assume that two different representational systems (episodic and abstract) co-exist in memory.

The experiments reported in this thesis also offer some tentative insight into variables that may mediate the contribution of these two systems to recognition and processing of lexical items. Experiment 5 suggests that the presence of pre-established representations in lexical memory may increase the speed with which there is a change in reliance between the the two systems, with the system underlying lexical competition effects becoming gradually more dominant over time. Experiment 7 further suggests that increased variability in the input enables new representations in the (presumably abstract) system underlying lexical competition effects to be stabilised more rapidly, resulting in an increase in the speed with which this system becomes dominant. This latter finding is consistent with research suggesting that increasing the amount of variability in the input enables the listener to home in on and extract the invariant (or 'abstract') properties of the speech signal (*e.g.*, Rost & McMurray, 2009).

6.2 A complementary learning systems account

As outlined in Chapter 2, the CLS framework (McClelland et al., 1995) may offer one account of a hybrid model, in which episodic and abstract representations are dependent on the hippocampal and neocortical subsystems respectively. Evidence suggests that new representations are generated rapidly in the hippocampal network (Breitenstein et al., 2005; Heckers et al., 2002; Mestres-Misse et al., 2008). Our results support this claim by showing that newly learned words are recognised

accurately immediately after study. These rapidly generated hippocampal representations are assumed to be highly detailed in nature, consistent with the finding that TSEs in recognition memory were present immediately after study in all experiments reported in this thesis. Information in the hippocampal subsystem is also assumed to be represented in a sparsely coded, pattern-separated manner that allows highly similar representations to be kept separate from one another (*e.g.,* Bakker et al., 2008). Evidence supporting this latter claim comes from Experiment 7 where significant TSEs were observed in the Day 1 test session despite the fact that each novel nonword was heard in two different voices during study. If episodic traces acquired during study were not pattern-separated then traces of the same novel nonword should overlap due to shared phonological content. If this were the case then TSEs should not have been observed on Day 1 following between-talker variability during study.

Previous research has revealed temporally graded retrograde amnesia spanning several years in patients with hippocampal damage (*e.g.,* Nadel & Moscovitch, 1997) suggesting that hippocampal representations are maintained for a considerable period of time post-acquisition. Although on a much shorter time-scale, the maintenance of TSEs in recognition memory over the course of a week in Experiments 2 and 3 is consistent with this suggestion. However, data from the male/female categorisation task in Experiments 3, 5, and 6 suggest that the ability to explicitly recall information about the study talker of a novel word decreases over time, indicating that the episodic traces gradually decay. Given this pattern of male/female categorisation data it seems likely that memory for the novel nonwords would eventually become independent of the hippocampal system and that TSEs in recognition memory would eventually decrease. This suggestion is in line with the proposal that there is a gradual change in reliance between the hippocampal and neocortical subsystems for 'semantic memories', consistent with both standard consolidation theories (Squire & Alvarez, 1995; Squire & Zola, 1998; Teng & Squire, 1999; Zola-Morgan & Squire, 1990) and MTT (Winocur & Moscovitch, 2011). Further evidence supporting this suggestion comes from the confidence ratings data, which indicate that the contribution of recollection to recognition memory decreases over time whilst familiarity processes become more important. As stated above, evidence indicating that these two processes rely on the hippocampus and surrounding MTL respectively is consistent with the CLS framework in which

there is a gradual loss of reliance on the highly-detailed hippocampally-mediated episodic representations over time. Neuroimaging studies showing that contributions from hippocampal regions to memory decrease over time whilst contributions from neocortical regions increase (*e.g.*, Bontempi et al., 1999; Takashima et al., 2009) further support this suggestion.

In comparison to the hippocampal subsystem, learning is assumed to occur more slowly in the neocortical subsystem, with multiple repetitions and/or periods of sleep-associated offline consolidation being required in order for a new representation to become robustly specified (McClelland et al., 1995). The combined analysis of lexical decision data in which lexical competition effects were absent immediately after study but emerged one day later, following a period of sleep-associated offline consolidation, is consistent with this suggestion. These slowly-acquired neocortical representations are thought to be distributed and overlapping in nature, enabling phonologically-similar words to compete during spoken word recognition, as observed in the combined analysis of lexical decision data on Day 2. As a result of this distributed, overlapping nature of representation it may be assumed that the neocortical system consists of a set of abstract representations. As noted above, the presence of lexical competition effects for both same- and different-talker items on Day 2 of the combined analysis supports this suggestion. However, the Day 8 data are more problematic. On the one hand the possibility that the talker-specific lexical competition effects observed on Day 8 may be explained in terms of co-activation of abstract and episodic representations of the novel nonwords is consistent with a hybrid CLS model and the notion that the hippocampal and neocortical systems are linked and highly interactive. This explanation would propose that abstract representations are strengthened during periods of sleep-associated offline consolidation and are mediated by neocortical regions, but can be activated in conjunction with hippocampally mediated episodic representations depending on the task demands and/or the amount of processing time. On the other hand, if talker information is consolidated alongside phonological information, with talker-specific lexical competition effects being dependent on this consolidated talker-specific information then this explanation poses more of a challenge to a hybrid CLS model of lexical representation. Specifically, this explanation is inconsistent with the suggestion that representations in the neocortical system are abstract in nature. Possible methods of differentiating between these two

explanations and clarifying the nature of representation in the neocortical subsystem of the CLS framework are outlined below in section *6.3.3*.

6.3 Limitations and future directions

Whilst the experiments reported in this thesis offer some insight into how new words are represented and how reliance on different types of representations change over time there are a number of confounds and additional questions that are yet to be addressed.

6.3.1 The time-course of talker-specificity effects for existing words

Firstly, the results of Experiment 5 hint at potentially important differences in the processing of existing and novel words in the delayed test session on Day 8. However, the lack of significant interactions in the main analysis makes it difficult to draw any clear conclusions from these data. It is possible that the relatively poor performance in Experiments 4 and 5 may have masked differences between processing of existing and novel words. In particular, the use of morphologically and pseudo-morphologically related targets and foils appears to have made the task very difficult, as evidenced by the much lower accuracy scores in old/new categorisation task compared to Experiments 1-3 and 6-7. Decreasing the number of items to be learned, and increasing the number of repetitions of each item during study should increase performance in the old/new categorisation task as well as possibly decreasing the amount of variability in performance between participants, potentially allowing subtle differences between existing and novel words to be observed more clearly. On the other hand, these changes may have the opposite effect, with an increased number of exposures to each item during study simply strengthening TSEs for all items at all time points. This latter finding would suggest that the different time-courses of TSEs for existing (Goldinger, 1996) and novel words (Experiments 2-3) observed in previous experiments are, after all, simply be due to the number of items studied and the number of exposures to each item prior to the test session.

An alternative way of equating the sets of existing and novel words may be to teach participants to associate the novel nonwords with novel semantic concepts or objects. However, studies by Tamminen (2010) suggest that novel semantic information associated with newly-learned words requires a considerable period of time in order to be fully integrated into the existing lexicon. Thus, it seems unlikely

that semantic information could be accessed and used in the same way to support recognition of existing and novel words immediately after study if this design were used. However, given that we are particularly interested in potential differences between processing and representation of existing and novel words primarily in the one-week retest the slow integration of novel semantic information into the existing lexicon may pose less of a problem.

6.3.2 The effects of surface-form variability on word learning

As with the comparison of TSEs for existing and novel words, the findings from Experiment 7 are not as clear as we would have liked them to be. Specifically, the pattern of data across the three test sessions indicated that TSEs were significant only in the immediate test session, not in the two delayed test sessions. Nonetheless, the interaction between test-phase talker and day was not significant in the main analysis. As suggested in Chapter 5 one way to clarify the pattern of TSEs following between-talker variability during study may be to introduce more talkers during the study phase of the experiment. If our predictions are correct then immediately after study, when the episodic subsystem is dominant, TSEs should be observed regardless of the number of talkers associated with each novel word during study due to pattern-separation of representations in the episodic subsystem. However, if variability is important for driving the formation of more robust abstract representations, then increasing the number of talkers should further decrease the size of TSEs in the delayed test sessions, and should, in theory, enable a significant interaction between test-phase talker and day to be observed.

An additional interesting design to consider would be to group the study talkers such that half of the items were heard only in female voices, and the other half only in male voices. This design would be useful in determining whether the stability of TSEs observed in Experiments 1-6 was due to the fact that only two talkers of different genders were used. If this were the case then TSEs should remain stable and close to significance at all time-points despite the presence of between-talker but within-gender variability. A further advantage of using this design is that it would be possible to re-introduce the male/female categorisation task in order to examine whether participants are still able to explicitly access information about the study talker of each novel word despite hearing each item spoken by more than one talker during study.

Finally, it is important to consider that the number of exposures to each individual talker during the study phase of Experiment 7 was half that used in Experiments 1, 2, 3, and 6. Whilst we have argued that the number of tokens per study talker is unlikely to be causally related to the pattern of data observed in Experiment 7 since TSEs in recognition memory were still significant on Day 1 despite the smaller number of exposures to each talker during study, it would be reassuring to demonstrate that the same pattern of data, specifically the absence of TSEs at delayed test points, was observed when there were 18 unique tokens spoken by each of two (or more) study talkers. Of course, this design itself has limitations in that this manipulation would double (or triple/quadruple *etc.*) the number of exposures to the phonological form of the novel nonwords during study. However, in conjunction with Experiment 7 and the experiments suggested above, data from this experiment would help to provide a more comprehensive (and hopefully clearer) set of data examining the effects of between-talker variability during study than is currently available in Chapter 5.

6.3.3 Lexical competition effects

With regards to the lexical competition data, if as suggested in Chapter 5 the lack of significant lexical competition effects in most of the experiments reported in this thesis is in fact due to a lack of power, then using a design in which participants learn a greater number of novel nonwords, and are exposed to these items a greater number of times during study (in order to strengthen memory for the phonological form of the novel items) should result in more robust lexical competition effects on Days 2 and 8. It would also be advisable to revert back to the design used in the lexical decision task in Experiment 2 where all participants heard both talkers during the test of lexical competition rather than using only one talker, as was the case in Experiments 3, 6, and 7. It is possible that when only one voice is used at test talker information is rendered less informative as a cue to spoken word recognition, and so participants place more emphasis on abstract lexical information contained in the speech signal (Goh, 2005). These changes to the design of the lexical competition task should strengthen and clarify the tentative pattern of talker-specific lexical competition effects observed in the experiments reported here.

Alternatively it might be useful to consider using eye-tracking to examine lexical competition effects between existing and novel words. An eye-tracking

design would offer greater insight into the time-point at which talker information begins to affect lexical competition between similar-sounding words. Using eye-tracking to examine the point at which talker information begins to influence spoken word recognition when exposed to pairs of phonologically-similar existing and novel words would also address Luce et al.'s (2003) time-course hypothesis, which predicts that talker information is only integrated if there is sufficient processing time. If this is the case then differences between same- and different-talker items should only be observed at late time-windows in an eye-tracking task. Creel and colleagues (Creel et al., 2008; Creel & Tumlin, 2009) have already examined the online time-course of TSEs in lexical competition between pairs of phonologically-similar existing words, and between pairs of phonologically-similar novel words. However, whilst there are a large number of eye-tracking studies examining lexical competition effects between pairs of existing words or pairs of novel words few studies have attempted to demonstrate integration of existing and novel words within an eye-tracking paradigm, and those that have provide limited evidence that existing and artificial lexicons interact (Magnuson, Tanenhaus, Aslin, & Dahan, 2003, Experiment 3). Thus it would be important to establish a robust paradigm using a single talker before introducing multiple talkers into the study and test-phase of an eye-tracking experiment.

One final point that needs to be addressed with regards to the lexical competition data is whether the talker-specific lexical competition effects on Day 8 are driven by a combination of hippocampal and neocortical representations that are episodic and abstract respectively, or whether talker information is consolidated and stored in the neocortical system itself, alongside phonological information. If both phonological and talker information are consolidated in memory, but different hemispheres mediate processing of these two types of information (*cf.* Marsolek, 1999) then behavioural studies using dichotic listening would predict larger talker-specific lexical competition effects when items are presented to the left ear, and thus contralaterally to the right hemisphere. Given the pattern of lexical competition effects observed in the experiments reported in this thesis it would be predicted that hemispheric asymmetries should only be observed on Day 8, there should be no hemispheric asymmetries when participants are tested on Day 2. Further evidence from an fMRI study may also provide further insights into whether Marsolek's (1999) dissociable neural subsystems approach is viable, or whether the simpler

account, in which talker-specific lexical competition effects arise from a combination of activation of hippocampal and neocortical representations of a novel word is more likely. If talker-specific lexical competition effects arise due to co-activation of abstract and episodic representations of novel words in neocortical and hippocampal networks respectively then greater hippocampal activation would be predicted on same-talker compared to different-talker trials in the Day 8 test session, whilst no difference in activation in the neocortical subsystem would be predicted. In comparison, if talker-specific lexical competition arises due to consolidation of talker information into the neocortical subsystem then greater neocortical activation might be expected on same-talker compared to different talker trials, particularly in RH regions. According to this latter explanation no difference in hippocampal activation would be predicted for same- and different-talker trials in the Day 8 test session. Thus, by comparing the patterns of activation on same- and different-talker lexical decision trials it should be possible to differentiate between these two explanations.

6.3.4 Generalisability of findings

Finally, given that all of the experiments reported in this thesis, and indeed all of the experiments suggested above, examine episodic effects for only one type of indexical information (talker identity), it would be interesting to examine whether the same time-course of surface-form specificity effects is observed in recognition memory and lexical competition effects for other types of indexical and/or allophonic variables such as speech rate and voice onset time. Similar patterns of data across multiple indexical (and possibly allophonic) variables would provide further support for a hybrid CLS model of lexical representation for newly-learned words. However, it may be that only indexical variables that are of communicative value (*e.g.,* pitch) and/or are likely to affect the phonetic properties of a word (*e.g.,* speech rate) that are encoded and stored in memory, and are able to influence later recognition of newly learned words; indexical variables that do not influence the phonetic realisation of a word (*e.g.,* amplitude) and/or are not of communicative value may not affect later recognition of recently learned words (*cf.* Bradlow et al., 1999).

6.4 Concluding remarks

To conclude, it has been proposed that new words are initially represented in a temporary store that contains highly detailed episodic representations. Within a CLS framework (McClelland et al., 1995) this corresponds to the hippocampal network. Representations in this episodic subsystem are maintained to some degree for at least a week after the new words are initially learned, although it appears that explicit access to these representations decreases over time. Following a period of sleep-associated offline consolidation representations in an abstract subsystem (the neocortical network in a CLS model) are strengthened, allowing new words to begin to engage in competition with similar sounding existing words. Over time, information about the study talker of the new word may also be consolidated, possibly forming links with the newly-generated abstract representations, allowing talker-specific lexical competition effects to emerge one week after the new words have been learned. Variables that may influence the speed with which there is a change in reliance between the episodic and abstract subsystems, at least in terms of which representations dominate during tests of recognition memory, include the novelty of an item and the amount of variability in the study tokens of the novel words. However, further investigation of these two variables is required before we can make stronger claims about their importance.

APPENDIX A

Base words, novel nonwords, and foil words used in Experiments 1, 2, 3, 6, and 7

Base-word	*Novel-word*	*Foil-word*	*Phonemes*	*CelexFreq*
amulet	amulos	amulok	9	2
anecdote	anecdel	anecden	9	3
artichoke	artiched	artichen	8	3
assassin	assassool	assassood	8	3
baboon	babeel	babeen	6	4
bayonet	bayoniss	bayonil	8	3
blossom	blossail	blossain	7	2
bramble	brambooce	bramboof	7	2
capsule	capsyod	capsyoff	8	5
caravan	caravoth	caravol	9	3
cataract	catarist	catarill	10	3
cathedral	cathedruke	cathedruce	10	3
clarinet	clarinern	clarinerl	10	3
consensus	consensom	consensog	11	14
daffodil	daffadat	daffadan	9	3
decibel	decibit	decibice	9	2
dolphin	dolpheg	dolphess	7	3
dungeon	dungeill	dungeic	7	2
gimmick	gimmon	gimmod	6	3
grimace	grimin	grimib	7	4
haddock	haddale	haddan	6	2
hormone	hormike	hormice	6	7
hurricane	hurricarb	hurricarth	9	3
hyacinth	hyasel	hyased	8	3
lantern	lantobe	lantoke	7	2
lectern	lectas	lectack	7	2
methanol	methanack	methanat	9	2
molecule	molekyen	molekyek	10	3
moped	mopall	mopass	6	2
mucus	muckip	muckin	7	3
octopus	octopoth	octopol	9	2
ornament	ornameast	ornameab	9	3
parachute	parasheff	parashen	9	3
parsnip	parsneg	parsnes	7	2
partridge	partred	partren	7	10
pedestal	pedestoke	pedestode	9	3
pelican	pelikiyve	pelikibe	9	3
profile	profon	profod	7	12
pulpit	pulpen	pulpek	7	5
pyramid	pyramon	pyramotch	9	3
siren	siridge	sirit	8	5
skeleton	skeletobe	skeletope	9	3
slogan	slowgiss	slowgith	7	2
spasm	spaset	spasel	7	5
specimen	specimal	specimav	10	3
squirrel	squirrome	squirrope	7	2
tavern	tavite	tavile	6	5
tycoon	tycol	tycoff	6	4

APPENDIX B

Effects of study and test voice (male *vs.* female) in tasks examining talker-specificity effects

For all tasks examining TSEs two additional sets of analyses were carried out. The first examined whether the data were affected by whether an item had been *studied* in the male or female voice. The second examined whether data were affected by whether an item had been *tested* in the male or female voice. These analyses were carried out in order to establish that the TSEs reported in Chapters 3-5 were not simply due to differences in which voice was heard at study and/or test, but were instead due to hearing the same or a different talker, and thus could be attributed to memory for detailed information about the study voice of an item.

Data were analysed in repeated-measures ANOVAs with the same factors as reported in the main analyses from Chapters 3-5, but with test-phase talker (same *vs.* different) replaced by study/test voice (male *vs.* female). Only significant main effects and interactions involving this variable are reported here.

Experiment 1

In the *stem completion* task study voice affected responses, with items studied in the male voice producing slower responses (male = 1446 ms; female = 1310 ms), $F_1(1,27) = 4.39$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,34) = 2.35$, *ns*, but greater accuracy (male = 52.1%, female = 44.5%), $F_1(1,28) = 8.37$, $p < .01$, $\eta_p^2 = .23$, $F_2(1,46) = 3.49$, $p = .068$, $\eta_p^2 = .07$. Test voice did not influence performance.

There were also main effects of study voice in the *delayed shadowing* task, with items shadowed significantly more accurately if they were spoken by the male talker at study (male = 87.5%, female = 83.0%), $F_1(1,27) = 18.42$, $p < .001$, $\eta_p^2 = .41$, $F_2(1,92) = 22.64$, $p < .001$, $\eta_p^2 = .20$. The main effect of test voice was significant only in the by-items analysis (male = 89.4%, female = 81.1%), $F_1(1,27) = 2.62$, *ns*, $F_2(1,92) = 9.38$, $p < .01$, $\eta_p^2 = .09$.

Neither d' nor β values in the *old/new categorisation* task differed depending on the study or test voice of the items.

Experiment 2

As in Experiment 1, neither d′ or β values in the *old/new categorisation* task differed depending on whether the study or test voice of the item was male or female

Experiment 3

Analysis of d′ data from the *old/new categorisation* task revealed significant main effects of both study voice, $F(1,46) = 12.52$, $p = .001$, $\eta_p^2 = .21$, and test voice, $F(2,92) = 4.81$, $p < .05$, $\eta_p^2 = .10$. Likewise, there were significant main effects of study voice, $F(1,46) = 7.56$, $p < .01$, $\eta_p^2 = .14$, and test voice, $F(1,46) = 11.27$, $p < .01$, $\eta_p^2 = .20$, for β values.

For d′ values categorisation of items studied in the female voice ($M = 1.93$) was poorer than categorisation of items originally studied in the male voice ($M = 2.19$). Consistent with this, with β values indicating that participants were more biased to classify items studied in the female voice as new (higher β, $M = 1.45$) than they were for male items (lower β, $M = 1.14$). Both of these findings may simply reflect the fact that more phoneme monitoring errors were produced for items heard in the female voice, indicating poorer encoding of these items during study.

However, items heard in the male voice at test were categorised more poorly (male = 1.96, female = 2.14), and participants were more biased to classify items heard in the male voice at test as new (male = 1.44, female = 1.09). These findings are difficult to account for. Nonetheless, given the similarity between Experiments 2 and 3 for both d′ and β data in the main analyses reported in Chapter 3 it seems unlikely that differences in responding to items heard in a different voice at test are able to account for the pattern of data, particularly since no effects of either study or test voice were found in the SDT analysis in Experiment 2.

Analysis of *AUC* data revealed that there were main effects of both study voice, $F(1,46) = 5.17$, $p < .05$, $\eta_p^2 = .10$, and test voice, $F(1,46) = 8.13$, $p < .01$, $\eta_p^2 = .15$, with higher AUC values for items studied in the male voice ($M = .93$) than items studied in the female voice ($M = .91$), but higher AUC values for items tested in the female voice ($M = .93$) compared to items tested in the male voice ($M = .90$), consistent with the data from the old/new categorisation task. Importantly, neither study or test voice interacted with day.

Accuracy in the *male/female categorisation* tasks did not differ depending on whether the study voice was male or female. However, accuracy did differ depending on whether the items were heard in the male voice or female voice at test, $F_1(1,46) = 6.82$, $p < .05$, $\eta_p^2 = .13$, $F_2(1,46) = 1.78$, *ns*, with more errors being made to items heard in the male voice at test ($M = 35.1\%$) than items heard in the female voice at test ($M = 31.7\%$). Nonetheless, it is important to note that this main effect was not significant in the by-items analysis, making the finding difficult to interpret.

Experiment 4

Analysis of the *old/new categorisation* data revealed that the main effect of study voice was not significant in either d′ or β analyses. However, the main effect of test voice was significant for the d′ data, $F(1,30) = 5.72$, $p < .02$, $\eta_p^2 = .16$, with participants showing greater accuracy for items heard in the male voice at test ($M = .72$) compared to items heard in the female voice at test ($M = .42$). Nevertheless, test voice did not interact with word-type, $F(1,30) = 1.83$, *ns*, suggesting that any differences observed between existing and novel words in the analyses reported in Chapter 4 cannot be accounted for by a male-advantage at test.

In the *AUC* analysis performance did not differ depending on the study or test voice of the items. Likwise, accuracy in the *male/female categorisation* task did not differ depending on the study or test voice of each item.

Experiment 5

Neither d′ nor β scores differed in the *old/new categorisation task* depending on the study or test voice (male *vs.* female) of each item.

In the *AUC* analysis the main effect of study voice and test voice were both non-significant. However, the three-way interaction between study voice, day, and word-type was marginally significant, $F(1,38) = 3.61$, $p = .065$, $\eta_p^2 = .09$, as were the two-way interactions between test-voice and day, $F(1,38) = 3.17$, $p = .083$, $\eta_p^2 = .08$, and test voice and word-type, $F(1,38) = 3.08$, $p = .087$, $\eta_p^2 = .08$. None of the remaining interactions were significant.

Accuracy in the *male/female categorisation* task did not differ overall depending on study voice. However, the interaction between study voice and day was marginally significant, $F_1(1,38) = 3.20$, $p = .082$, $\eta_p^2 = .08$, $F_2(1,92) = 3.85$, $p =$

.053, η_p^2 = .04. Nonetheless, study voice did not interact with word-type, suggesting that any differences observed between existing and novel words, reported in Chapter 4, cannot be accounted for by the study voice of the items affecting existing and novel words differently. There was also a significant main effect of test voice in the *male/female categorisation* data, $F_1(1,38) = 4.12$, $p < .05$, $\eta_p^2 = .10$, $F_2(1,92) = 3.03$, $p = .085$, $\eta_p^2 = .03$, as well as a significant interaction between test voice, day, and word-type, $F_1(1,38) = 6.57$, $p < .05$, $\eta_p^2 = .15$, $F_2(1,92) = 6.92$, $p = .01$, $\eta_p^2 = .07$. Analysis of each word-type separately on each day revealed that this was due to a significant main effect of test voice only for existing words on Day 1, $F(1,38) = 5.94$, $p < .05$, $\eta_p^2 = .14$, $F(1,46) = 4.18$, $p < .05$, $\eta_p^2 = .08$, but only for novel words on Day 8, $F(1,38) = 5.08$, $p < .05$, $\eta_p^2 = .12$, $F(1,46) = 5.22$, $p < .05$, $\eta_p^2 = .10$, both revealing more errors for items heard in the male voice at test.

Experiment 6

No significant main effects of study and test voice, or interactions involving these variables, were found in any of the measures in Experiment 6 (d′, β, AUC, male/female categorisation)

Experiment 7

Likewise, no significant main effects of study and test voice, or interactions with these variables, were found in Experiment 7 (d′, β, AUC).

APPENDIX C

Analysis of response time data from the old/new categorisation task

Experiment 1

Response latencies from the old/new categorisation task in Experiment 1 were analysed using a repeated-measures ANOVA with factors test-phase talker (same *vs.* different to study), word-type (novel *vs.* foil), and task-order (stem completion first *vs.* old/new categorisation with delayed shadowing first). Analysis revealed a significant main effect of test-phase talker, $F_1(1,27) = 32.93$, $p < .001$, $\eta_p^2 = .55$, $F_2(1,84) = 20.54$, $p < .001$, $\eta_p^2 = .20$, with RTs to items heard in a different voice to study being longer than those heard in the same voice. The main effects of word-type and task-order were both non-significant (word type – $F_1(1,27) = 1.71$, *ns*, $F_2(1,84) = 1.58$, *ns*; task order – $F_1(1,27) = .12$, *ns*, $F_2(1,84) = .32$, *ns*). The only significant interaction was between test-phase talker and word-type, $F_1(1,27) = 10.12$, $p < .01$, $\eta_p^2 = .27$, $F_2(1,84) = 4.84$, $p = .05$, $\eta_p^2 = .05$ (Table C1). Post hoc comparisons revealed that the difference between same and different test-phase talker was significant for both novel nonwords, $F_1(1,27) = 27.36$, $p < .001$, $\eta_p^2 = .50$, $F_2(1,40) = 17.86$, $p < .001$, $\eta_p^2 = .31$, and foil nonwords, $F_1(1,27) = 4.40$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,44) = 3.54$, $p = .067$, $\eta_p^2 = .07$, although the effect size was larger for novel compared to foil nonwords.

Table C1. RTs (ms) to novel nonwords and foil nonwords in the old/new categorisation task (Experiments 1, 2, 3, 6, and 7) when spoken in the same or a different voice to study.

		Novel word		Foil word	
Exp	**Day**	Same	Different	Same	Different
1	1	1341	1496	1428	1468
2	1	1182	1265	1284	1263
	2	1101	1163	1227	1204
	8	1122	1184	1198	1225
3	1	1870	2111	1995	1921
	2	1672	1849	1831	1825
	8	1615	1790	1741	1746
6	1	2217	2345	2279	2226
	2	2001	2127	2088	2064
	8	1960	2097	1960	1951
7	1	1781	1851	1930	1959
	2	1636	1693	1791	1862
	8	1638	1641	1755	1723

Experiment 2

RT data from the old/new categorisation task in Experiment 2 (Table C1) were analysed using a repeated-measures ANOVA with factors test-phase talker (same *vs.* different), word-type (novel *vs.* foil), and day (1, 2, *vs.* 8). There was a significant main effect of word-type, $F_1(1,29) = 16.79$, $p < .001$, $\eta_p^2 = .37$, $F_2(1,88) = 10.37$, $p < .01$, $\eta_p^2 = .11$, indicating that RTs were quicker overall for novel nonwords compared to foil nonwords. This contrasts with Experiment 1 where there was no significant difference between RTs to novel and foil nonwords. There was also a significant main effect of day, $F_1(2,58) = 5.32$, $p < .01$, $\eta_p^2 = .16$, $F_2(1.7,146.5) = 19.24$, $p < .001$, $\eta_p^2 = .18$, most likely due to task-repetition across sessions. More importantly, there was a significant main effect of test-phase talker, $F_1(1,29) = 6.26$, $p < .05$, $\eta_p^2 = .18$, $F_2(1,88) = 5.73$, $p < .05$, $\eta_p^2 = .06$, as well as an interaction between word-type and test-phase talker, $F_1(1,29) = 11.23$, $p < .01$, $\eta_p^2 = .28$, $F_2(1,88) = 6.69$, $p < .05$, $\eta_p^2 = .07$. Separate analysis for novel and foil nonwords indicated that there were significant main effects of test-phase talker only for novel words, $F_1(1,29) = 15.22$, $p = .001$, $\eta_p^2 = .34$, $F_2(1,45) = 12.21$, $p = .001$, $\eta_p^2 = .21$, with same-talker items being responded to quicker than different-talker items. The null effect for the foil nonwords indicated that RTs to these items did not differ as a function of whether the foil word was heard in the same or a different voice to that in which its corresponding novel nonword had been studied. Importantly, the interaction between test-phase talker and day was non-significant for both novel and foil words, as in the main analysis, suggesting that the size of the TSEs did not change across test-sessions.

Experiment 3

RT data from Experiment 3 (Table C1) were analysed using a repeated-measures ANOVA with the same factors as Experiment2. Analysis revealed a significant main effect of test-phase talker, $F_1(1,46) = 42.18$, $p < .001$, $\eta_p^2 = .48$, $F_2(1,89) = 26.40$, $p < .001$, $\eta_p^2 = .23$, with faster RTs to same-talker items. There was also a significant main effect of day, $F_1(2,92) = 19.22$, $p < .001$, $\eta_p^2 = .39$, $F_2(1.8,162.9) = 116.71$, $p < .001$, $\eta_p^2 = .57$, likely reflecting practice effects due to task repetition across the three test points. However, contrary to Experiment 2 the main effect of word-type (novel *vs.* foil) was not significant, $F_1(1,46) = 2.37$, *ns,*

$F_2(1,89) = .91$, *ns*, although the interaction between word-type and test-phase talker was significant, $F_1(1,46) = 55.61$, $p < .001$, $\eta_p^2 = .55$, $F_2(1,89) = 43.62$, $p < .001$, $\eta_p^2 = .33$. In the by-participants analysis the two-way interaction between word-type and day was also significant, $F_1(2,92) = 6.15$, $p < .01$, $\eta_p^2 = .12$, as was the three-way interaction between test-phase talker, word-type, and day, $F_1(2,92) = 3.39$, $p < .05$, $\eta_p^2 = .07$, although neither of these were significant by-items (word-type x day - $F_2(2.178) = 2.36$, *ns*; test-talker x day x word-type - $F_2(1,178) = 2.03$, *ns*). Separate analysis for novel words and foil words replicated findings from Experiment 2, with a significant main effects of test-phase talker for novel words, $F_1(1,46) = 100.08$, $p < .001$, $\eta_p^2 = .69$, $F_2(1,45) = 48.58$, $p < .001$, $\eta_p^2 = .52$, but not foil words, $F_1(1,46) = 1.53$, *ns*, $F_2(1,45) = 1.92$, *ns*. Likewise, the interaction between test-phase talker and day was not significant for either novel or foil nonwords, suggesting once again that the size of the TSEs did not change across test-sessions.

Experiment 4

Analysis of RT data from the old/new categorisation task in a repeated-measures ANOVA, with factors test-phase talker (same *vs.* different), word-type (existing *vs.* novel), and studied versus unstudied, revealed a significant main effect of word-type, $F_1(1,30) = 4.86$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,82) = 6.14$, $p < .05$, $\eta_p^2 = .07$, with faster RTs to novel than existing words (Table C2). Studied items were also responded to faster than unstudied items, $F_1(1,30) = 18.44$, $p < .001$, $\eta_p^2 = .38$, $F_2(1,82) = 15.89$, $p < .001$, $\eta_p^2 = .16$, and the main effect of test-phase talker was marginally significant, $F_1(1,30) = 3.28$, $p = .080$, $\eta_p^2 = .10$, $F_2(1,82) = 3.60$, $p = .061$, $\eta_p^2 = .04$. Given that there was a significant interaction in the by-participants analysis between test-phase talker and whether an item had been studied or not, $F_1(1,30) = 5.78$, $p < .05$, $\eta_p^2 = .16$, RT data were analysed separately for studied and unstudied items. For studied items there was a significant main effect of test-phase talker in the RT data, $F_1(1,30) = 10.04$, $p < .01$, $\eta_p^2 = .25$, $F_2(1,92) = 3.20$, $p = .077$, $\eta_p^2 = .03$, with faster RTs to same-talker items. This same-talker advantage did not extend to the unstudied items, $F_1(1,30) = .38$, *ns*, $F_2(1,82) = .59$, *ns*. This latter finding is consistent with Experiment 3, in which the main effect of test-phase talker was non-significant for foil nonwords.

Experiment 5

Analysis of RT data using the same repeated-measures ANOVA as in Experiment 4 but with the additional factor day (1, 2, *vs.*8), revealed significant main effects of day, $F_1(1,36) = 4.59$, $p < .05$, $\eta_p^2 = .11$, $F_2(1,87) = 60.22$, $p < .001$, $\eta_p^2 = .41$, word-type, $F_1(1,36) = 14.79$, $p < .001$, $\eta_p^2 = .29$, $F_2(1,87) = 18.14$, $p < .001$, $\eta_p^2 = .17$, and studied/unstudied, $F_1(1,36) = 52.08$, $p < .001$, $\eta_p^2 = .59$, $F_2(1,87) = 51.46$, $p < .001$, $\eta_p^2 = .37$, indicating faster RTs on Day 8, for novel words, and for studied items respectively (Table C2). The main effect of test-phase talker was not significant, $F_1(1,36) = 2.47$, *ns*, $F_2(1,87) = 2.39$, *ns*. However, there was a significant interaction between test-phase talker and studied/unstudied, $F_1(1,36) = 10.71$, $p < .01$, $\eta_p^2 = .23$, $F_2(1,87) = 4.28$, $p < .05$, $\eta_p^2 = .05$. Additionally, the three-way interaction between word-type, test-phase talker, and studied/unstudied was marginally significant, $F_1(1,36) = 3.73$, $p = .061$, $\eta_p^2 = .09$, $F_2(1,87) = 1.27$, *ns*. Separate analysis of studied and unstudied items revealed a significant main effect of test-phase talker only for studied items, $F_1(1,37) = 12.75$, $p = .001$, $\eta_p^2 = .26$, $F_2(1,92) = 11.13$, $p = .001$, $\eta_p^2 = .11$, as well as an interaction between test-phase talker and word-type for studied items in the by-participants analysis, $F_1(1,37) = 5.34$, $p < .05$, $\eta_p^2 = .13$, (but not by-items $F_2(1,92) = 1.58$, *ns*). This interaction reflected the fact that the main effect of test-phase talker was significant for studied novel words, $F_1(1,37) = 22.13$, $p < .001$, $\eta_p^2 = .37$, $F_2(1,46) = 11.74$, $p = .001$, $\eta_p^2 = .20$, but not studied existing words, $F_1(1,37) = 1.22$, *ns*, $F_2(1,46) = 1.97$, *ns*. This was true for both Day 1 and Day 8 data.

Table C2. RTs (ms) to studied and unstudied existing and novel words in the old/new categorisation task (Experiments 4 and 5) when spoken in the same or a different voice to study.

			Existing word		Novel word	
Exp	**Day**		Same	Different	Same	Different
4	1	Studied	2212	2338	2163	2311
		Unstudied	2606	2527	2427	2446
5	1	Studied	2157	2203	2045	2238
		Unstudied	2534	2455	2305	2263
	8	Studied	2055	2103	1941	2074
		Unstudied	2276	2286	2185	2143

Experiment 6

Analysis of RT data from the old/new categorisation task using a repeated-measures ANOVA identical to that used in Experiments 2 and 3 revealed significant main effects of test-phase talker, $F_1(1,29) = 4.63$, $p < .05$, $\eta_p^2 = .14$, $F_2(1,84) = 9.81$, $p < .01$, $\eta_p^2 = .11$, and day, $F_1(2,58) = 15.99$, $p < .011$, $\eta_p^2 = .36$, $F_2(2,168) = 54.56$, $p < .001$, $\eta_p^2 = .39$, indicating that overall RTs were faster for same-talker items, and unsurprisingly that RTs decreased across test sessions, again most likely resulting from practice effects and task repetition. Although there was no difference in RTs to novel and foil items overall, $F_1(1,29) = .16$, *ns*, $F_2(1,84) = .00$, *ns*, there was a significant interaction between word-type and test-phase talker, $F_1(1,29) = 25.92$, $p < .001$, $\eta_p^2 = .47$, $F_2(1,84) = 12.28$, $p = .001$, $\eta_p^2 = .13$ (Table C1). Further analysis revealed that the main effect of test-phase talker was significant only for novel nonwords, $F_1(1,29) = 14.44$, $p = .001$, $\eta_p^2 = .33$, $F_2(1,42) = 31.16$, $p < .001$, $\eta_p^2 = .43$, not for unstudied foil items, $F_1(1,29) = .99$, *ns*, $F_2(1,42) = .05$, *ns*, consistent with Experiments 2-4. This main-effect of test-phase talker for studied items was significant (or marginally significance) at all time points (Day 1 – $F_1(1,29) = 20.61$, $p < .001$, $\eta_p^2 = .42$, $F_2(1,42) = 27.39$, $p < .001$, $\eta_p^2 = .40$; Day 2 – $F_1(1,29) = 3.41$, $p = .075$, $\eta_p^2 = .11$, $F_2(1,42) = 11.11$, $p < .001$, $\eta_p^2 = .21$; Day 8 – $F_1(1,29) = 5.47$, $p < .05$, $\eta_p^2 = .16$), $F_2(1,42) = 2.43$, *ns*. None of the other interactions approached significance.

Experiment 7

RTs from the old/new categorisation task in Experiment 7 (Table C1) revealed that there were significant main effects of word-type, $F_1(1,28) = 18.38$, $p < .001$, $\eta_p^2 = .40$, $F_2(1,90) = 17.96$, $p < .001$, $\eta_p^2 = .17$, and day, $F_1(2,56) = 10.61$, $p < .001$, $\eta_p^2 = .28$, $F_2(2,180) = 44.16$, $p < .001$, $\eta_p^2 = .33$, indicating that RTs were faster for novel compared to foil nonwords, and that RTs decreased across the course of the week, most likely due to task repetition and practice effects. However, the main effect of test-phase talker was not significant, $F_1(1,28) = 2.75$, *ns*, $F_2(1,90) = .10$, *ns*, nor did any of the interactions approach significance.

APPENDIX D

Existing and novel word-pairs used in Experiments 4 and 5

Existing words		**Novel words**	
List 1	*List 2*	*List 1*	*List 2*
baptism	baptist	anecdent	anecdence
coherent	coherence	assassant	assassance
colonise	colonise	badmintant	badmintance
cubist	cubism	bayonize	bayonist
defiant	defiance	cartrist	cartrism
finalize	finalist	catarize	catarist
fragrance	fragrant	clarinence	clarinent
hesitance	hesitant	culprent	culprence
idealist	idealize	decibist	decibize
indulgent	indulgence	dungism	dungist
negligent	negligence	gelatant	gelatance
nudist	nudism	gimmence	gimment
obedience	obedient	hurricance	hurricant
publicize	publicist	hyasize	hyasist
racism	racist	methanance	methanant
radiance	radiant	mucism	mucist
realism	realist	napkist	napkism
subservience	subservient	ornamist	ornamize
subsistent	subsistence	parashist	parashize
symbolize	symbolist	parsnism	parsnist
theorist	theorize	pedestance	pedestant
unionize	unionist	sirent	sirence
variance	variant	spasence	spasent
vigilant	vigilance	yogism	yogist

REFERENCES

Abutalebi, J., Keim, R., Brambati, S. M., Tettamanti, M., Cappa, S. F., de Bleser, R., et al. (2007). Late acquisition of literacy in a native language. *Human Brain Mapping, 28*, 19-33.

Addis, D. R., Moscovitch, M., Crawley, A. P., & McAndrews, M. P. (2004). Recollective qualities modulate hippocampal activation during autobiographical memory retrieval. *Hippocampus, 14*(6), 752-762.

Adlam, A. L. R., Malloy, M., Mishkin, M., & Vargha-Khadem, F. (2009). *Dissociation between recognition and recall in developmental amnesia.* Paper presented at the Meeting on Episodic Memory and the Brain, Tallinn, Estonia.

Aggleton, J. P., Vann, S. D., Denby, C., Dix, S., Mayes, A. R., Roberts, N., et al. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia, 43*(12), 1810-1823.

Altmann, G. T. M. (1997). *The ascent of Babel: An exploration of language, mind, and understanding*. Oxford: Oxford University Press.

Alvarez, P., Zola-Morgan, S., & Squire, L. R. (1995). Damage limited to the hippocampal region produces long-lasting memory impairment in monkeys. *Journal of Neuroscience, 15*(5), 3796-3807.

Arnon, I., & Ramscar, M. (2009). *Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned.* Paper presented at the Annual Meeting of the Cognitive Science Society, Amsterdam.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX Lexical Database. Philadelphia, PA: Linguistic Data Consortium, University of Philadelphia.

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (2008). Pattern separation in the human hippocampus CA3 and dentate gyrus. *Science, 319*, 1640-1642.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L. H. (1990). Comparison of spatial and temporal characteristics of neuronal-activity in sequential stages of hippocampal processing. *Progress in Brain Research, 83*, 287-300.

Bastin, C., Van der Linden, M., Charnallet, A., Denby, C., Montaldi, D., Roberts, N., et al. (2004). Dissociation between recall and recognition memory performance in an amnesic patient with hippocampal damage following carbon monoxide poisoning. *Neurocase, 10*(4), 330-344.

Begg, I. (1971). Recognition memory for sentence meaning and wording. *Journal of Verbal Learning and Verbal Behaviour, 10*, 176-181.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*(3), 389-395.

Bontempi, B., Laurent-Demir, C., Destrade, C., & Jaffard, R. (1999). Time-dependent reorganization of brain circuitry underlying long-term memory storage. *Nature, 400*(6745), 671-675.

Borovsky, A., Elman, J., & Kutas, M. (2010). *Semantic integration of novel word meanings after a single exposure in context.* Paper presented at the 32nd Annual Conference of the Cognitive Science Society.

Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition, 116*, 289-296.

Boss, B. D., Turlejski, K., Stanfield, B. B., & Cowan, W. M. (1987). On the number of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Winstar rats. *Brain Research, 406*(1-2), 280-287.

Bowers, J. S. (2000). In defense of abstractionist theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 7*(1), 83-99.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Interfering neighbours: The impact of novel word learning on the identification of visually similar words. *Cognition, 97*(3), B45-B54.

Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Priessner, J. C., et al. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 16382-16387.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977-985.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206-219.

Brandt, K. R., Gardiner, J. M., Vargha-Khadem, F., Baddeley, A. D., & Mishkin, M. (2009). Impairment of recollection but not familiarity in a case of developmental amnesia. *Neurocase, 15*(1), 60-65.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A. F., Sommer, J., Wolbers, T., et al. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *Neuroimage, 25*(3), 958-968.

Brown, H., Weighall, A., Henderson, L. M., & Gaskell, M. G. (in press). Enhanced recognition and recall of new words in 7- and 12-year old children following a period of offline consolidation. *Journal of Experimental Child Psychology.*

Brown, J. S., & Carr, T. H. (1993). Limits on perceptual abstraction in reading: Asymmetric transfer between surface forms differing in typicality. *Journal of Experimental Psychology-Learning Memory and Cognition, 19*(6), 1277-1296.

Burgund, E. D., & Marsolek, C. J. (1997). Letter-case-specific priming in the right cerebral hemisphere with a form-specific perceptual identification task. *Brain and Cognition, 35*(2), 239-258.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*(3), 256-284.

Capone, N. C., & McGregor, K. K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech Language and Hearing Research, 48*(6), 1468-1480.

Carey, S. (1978). The child as a word learner. In M. Halle, J. Bresnan & A. Miller (Eds.), *Linguistic Theory and Psychological Reality* (pp. 264-293). Cambridge, MA: MIT Press.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(3), 521-533.

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.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47*, 207-239.

Cousineau, D. (2007). Confidence intervals in within-subject designs: A simpler solution to Loftus & Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42-45.

Craik, F. I. M. (1991). On the specificity of procedural memory. In W. Kessen, A. Ortony & F. I. M. Craik (Eds.), *Memories, thoughts, and emotions: Essays in honor of George Mandler* (pp. 183-197). Hillsdale, NJ: Erlbaum.

Craik, F. I. M., & Kirsner, K. (1974). Effect of speakers voice on word recognition. *Quarterly Journal of Experimental Psychology, 26*, 274-284.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633-664.

Creel, S. C., & Tumlin, M. A. (2009). *Talker information is not normalized in fluent speech: Evidence from on-line processing of spoken words.* Paper presented at the Annual Meeting of the Cognitive Science Society, Amsterdam.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*, 264-285.

Cutler, A. (2008, Apr). *The 34th Sir Frederick Bartlett Lecture The abstract representations in speech processing.* Paper presented at the Annual Meeting of the Experimental-Psychology-Society, Birmingham, England.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2006). Coping with speaker-related variation via abstract phonemic categories. Proceedings of the 10th Conference on Laboratory Phonology, Paris.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*(5-6), 507-534.

Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2006). Triple dissociation in the medial temporal lobes: Recollection, familiarity, and novelty. *Journal of Neurophysiology, 96*(4), 1902-1911.

Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences of the United States of America, 100*(4), 2157-2162.

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*.

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*.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives; Perceptual learning of distorted speech:

Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology-General, 134*(2), 222-241.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences, 11*(9), 379-386.

Dockrell, J., Braisby, N., & Best, R. (2007). Children's acquisition of science terms: Simple exposure is insufficient. *Learning and Instruction, 17,* 577-594.

Dorfman, J. (1994). Sublexical components in implicit memory for novel words. *Journal of Experimental Psychology: Learniing, Memory, & Cognition, 20(5),* 1108-1125.

Duff, M. C., Hengst, J., Tranel, D., & Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature Neuroscience, 9*(1), 140-145.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*(1), 35-39.

Dumay, N., Gaskell, M. G., & Feng, X. (2004). *A day in the life of a spoken word*. Proceedings of the 26th Annual Conference of the Cognitive Science Society.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89*(6), 627-661.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30,* 123-152.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119*(4), 1950-1953.

Eldridge, L. L., Knowlton, B. T., Furmanski, C. S., Bookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: a selective role for the hippocampus during retrieval. *Nature Neuroscience, 3*(11), 1149-1152.

Elfman, K. W., Parks, C. M., & Yonelinas, A. P. (2008). Testing a neurocomputational model of recollection, familiarity, and source recogition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 752-768.

Ernestus, M. (2009). *The roles of reconstruction and lexcial storage in the comprehension of regular pronunciation variants.* Paper presented at the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton, UK.

Feustel, T. C., Shiffrin, R. M., & Salasoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology-General, 112*(3), 309-346.

Forster, J. C., & Forster, K. I. (2003). DMDX: A Windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, & Computers, 35*, 116-124.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature, 431*(7005), 188-191.

Frazier Norbury, C. F., Griffiths, H., & Nation, K. (2010). Sound before meaning: Word learning in autistic disorders. *Neuropsychologia, 48,* 4012-4019.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3*(4), 128-135.

Gagnepain, P., Henson, R., & Davis, M. (2010). *Learning and consolidation have different effects on the time-course of novel word recognition: An MEG*

study. Paper presented at the Architectures and Mechanisms in Language Processing, York, UK.

Gais, S., Albouy, G., Boly, M., Dang-Vu, T. T., Darsaud, A., Desseilles, M., et al. (2007). Sleep transforms the cerebral trace of declarative memories. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 18778-18783.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*, 105-132.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5-6), 613-656.

Geiselman, R. E., & Bellezza, F. S. (1977). Incidental retention of speaker's voice. *Memory & Cognition, 5*(6), 658-665.

Geiselman, R. E., & Crawley, J. M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior, 22*(1), 15-23.

Gibson, J. M., Brooks, J. O., Friedman, L., & Yesavage, J. A. (1993). Typography manipulations can affect priming of word stem completion in older and younger adults. *Psychology and Aging, 8*(4), 481-489.

Gilboa, A., Winocur, G., Grady, C. L., Hevenor, S. J., & Moscovitch, M. (2004). Remembering our past: Functional neuroanatomy of recollection of recent and very remote personal events. *Cerebral Cortex, 14*(11), 1214-1225.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1-67.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(1), 40-53.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology-Learning Memory and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the International Congress of Phonetic Sciences, 15*, 49-54.

Goldinger, S. D., Azuma, T., Kleider, H. M., & Holmes, V. M. (2003). Font-specific memory: More than meets the eye. In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 157-196). Oxford: Oxford University Press.

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition, 27*(2), 328-338.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology-Learning Memory and Cognition, 17*(1), 152-162.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Gonzalez, J., Cerveza-Crespo, T., & McLennan, C. T. (2010). Hemispheric differences in specificity effects in talker identification. *Attention, Perception, & Psychophysics, 72*(8).

Gonzalez, J., & McLennan, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology-Human Perception and Performance, 33*(2), 410-424.

Gonzalez, J., & McLennan, C. T. (2009). Hemispheric differences in the recognition of environmental sounds. *Psychological Science, 20*(7), 887-894.

Graf, P., & Ryan, L. (1990). Transfer-appropriate processing for implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*(6), 978-992.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition, 17*(5), 572-581.

Heckers, S., Weiss, A. P., Alpert, N. M., & Schacter, D. L. (2002). Hippocampal and brain stem activation during word retrieval after repeated and semantic encoding. *Cerebral Cortex, 12*(9), 900-907.

Henderson, L. M., Weighall, A., Brown, H., & Gaskell, M. G. (submitted-a). Consolidation of vocabulary is associated with sleep in children.

Henderson, L. M., Weighall, A., Brown, H., & Gaskell, M. G. (submitted-b). Online lexical competition during spoken word recognition and word learning in children and adults

Henson, R. N. A., Cansino, S., Herron, J. E., Robb, W. G. K., & Rugg, M. D. (2003). A familiarity signal in human anterior medial temporal cortex? *Hippocampus, 13*(2), 301-304.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551.

Hintzman, D. L., Block, R. A., & Inskeep, N. R. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 741-749.

Holdstock, J. S., Mayes, A. R., Gong, Q. Y., Roberts, N., & Kapur, N. (2005). Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. *Hippocampus, 15*(2), 203-215.

Horst, J. S., McMurray, B., & Samuelson, L. K. (2005). *Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture*. Paper presented at the Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society.

Horst, J. S., McMurray, B., & Samuelson, L. K. (2006). *Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture*. Paper presented at the Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology-Human Perception and Performance, 26*(5), 1570-1582.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology-Human Perception and Performance, 29*(6), 1143-1154.

Jacoby, L. L. (1983a). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology-Learning Memory and Cognition, 9*(1), 21-38.

Jacoby, L. L. (1983b). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior, 22*(5), 485-508.

Jacoby, L. L., & Hayman, C. A. G. (1987). Specific visual transfer in word identification. *Journal of Experimental Psychology-Learning Memory and Cognition, 13*(3), 456-463.

Joos, M. (1948). *Acoustic Phonetics*. Baltimore: Linguistic Society of America.

Ju, M., & Luce, P. A. (2006). Representational specificity of within-category phonetic variation in the long-term mental lexicon. *Journal of Experimental Psychology-Human Perception and Performance, 32*(1), 120-138.

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month old infants. *Cognition, 43*(3), 253-291.

Keenan, J. M., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behaviour, 16*, 549-560.

Kirsner, K. (1973). An analysis of the visual component in recognition memory for visual stimuli. *Memory & Cognition, 1*(4), 449-453.

Kirsner, K. (1974). Modality differences in recognition memory for words and their attributes. *Journal of Experimental Psychology, 102*(4), 579-584.

Kolers, P. A., & Ostry, D. J. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior, 13*(6), 599-612.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol Sci, 21*(11), 1551-1556.

Kouider, S., & Dupoux, E. (2005). Subliminal speech priming. *Psychological Science, 16*(8), 617-625.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141-178.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts - How listeners adjust to speaker variability. *Psychological Science, 19*(4), 332-338.

Lachs, L., McMichael, K., & Pisoni, D. B. (2003). Speech perception and implicit memory: Evidence for detailed episodic encoding in phonetic events. In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 215-235). Oxford: Oxford University Press.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55*(4), 306-353.

Lee, J. L. C. (2009). Reconsolidation: Maintaining memory relevance. *Trends in Neuroscience, 32*(8), 413-420.

Levitin, D. J., & Cook, P. R. (1996). Memory for musical tempo: Additional evidence that auditory memory is absolute. *Perception & Psychophysics, 58*(6), 927-935.

Light, L. L., Stansbury, C., Rubin, C., & Linde, S. (1973). Memory for modality of presentation: Within-modality discrimination. *Memory & Cognition, 1*(3), 395-400.

Lindsay, S., & Gaskell, M. G. (submitted). Lexical Integration of Novel Words using Spaced Learning and Testing.

Lindsay, S., Sedin, L., & Gaskell, M.G. (2012). Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. *Journal of Memory and Language, 66*(1), 210-225.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*(3), 1242-1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/ III: Long-term retentnion of new phonetic categories. *Journal of the Acoustical Society of America, 96*(4), 2076-2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89*(2), 874-886.

Lovelace, E. A., & Southall, S. D. (1983). Memory for words in prose and their locations on the page. *Memory & Cognition, 11*(5), 429-434.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26*(4), 708-715.

Luce, P. A., & Lyons, E. A. (1999). Processing lexically embedded spoken words. *Journal of Experimental Psychology-Human Perception and Performance, 25*(1), 174-183.

Luce, P. A., McLennan, C. T., & Charles-Luce, J. (2003). Abstractness and specificty in spoken-word recognition: indexical and allophonic variability in long-term repetition priming. In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking Implicit Memory* (pp. 197-214). Oxford: Oxford University Press.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology-Human Perception and Performance, 33*(2), 391-409.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology-General, 132*(2), 202-227.

Maguire, E. A., Vargha-Khadem, F., & Mishkin, M. (2001). The effects of bilateral hippocampal damage on fMRI regional activations and interactions during memory retrieval. *Brain, 124*, 1156-1170.

Manns, J. R., & Squire, L. R. (1999). Impaired recognition memory on the doors and people test after damage limited to the hippocampal region. *Hippocampus, 9*(5), 495-499.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 385*, 813-815.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London Series B-Biological Sciences, 176*(1043), 161-&.

Marr, D. (1971). Simple memory: Theory for archicortex. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 262*(841), 23-&.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*(1-2), 71-102.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*(4), 653-675.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology-Human Perception and Performance, 15*(3), 576-585.

Marsolek, C. J. (1999). Dissociable neural, subsystems underlie abstract and specific object recognition. *Psychological Science, 10*(2), 111-118.

Marsolek, C. J. (2003). What is priming and why? In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 41-64). Oxford: Oxford University Press.

Marsolek, C. J. (2004). Abstractionist versus exemplar-based theories of visual word priming: A subsystems resolution. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 57*(7), 1233-1259.

Marsolek, C. J., & Burgund, E. D. (2008). Dissociable neural subsystems underlie visual working memory for abstract categories and specific exemplars. *Cognitive Affective & Behavioral Neuroscience, 8*(1), 17-24.

Marsolek, C. J., Kosslyn, S. M., & Squire, L. R. (1992). Form-specific visual priming in the right cerebral hemisphere. *Journal of Experimental Psychology-Learning Memory and Cognition, 18*(3), 492-508.

Marsolek, C. J., Schacter, D. L., & Nicholas, C. D. (1996). Form-specific visual priming for new associations in the right cerebral hemisphere. *Memory & Cognition, 24*(5), 539-556.

Marsolek, C. J., Squire, L. R., Kosslyn, S. M., & Lulenski, M. E. (1994). Form-specific explicit and implicit memory in the right cerebral hemisphere. *Neuropsychology, 8*(4), 588-597.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology-Learning Memory and Cognition, 15*(4), 676-684.

Mattys, S. L., & Clark, J. H. (2002). Lexical activity in speech processing: evidence from pause detection. *Journal of Memory and Language, 47*(3), 343-359.

Mattys, S. L., & Liss, J. M. (2008). On building models of spoken-word recognition: Where there is as much to learn from natural "oddities" as artificial normality. *Perception & Psychophysics*.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543-562.

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. *Hippocampus, 12*(3), 325-340.

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Montaldi, D., Grigor, J., Gummer, A., et al. (2004). Associative recognition in a patient with selective hippocampal lesions and relatively normal item recognition. *Hippocampus, 14*(6), 763-784.

Mayes, A. R., & Montaldi, D. (2001). Exploring the neural bases of episodic and semantic memory: the role of structural and functional neuroimaging. *Neuroscience and Biobehavioral Reviews, 25*(6), 555-573.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

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.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159-188.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 24, pp. 109-164). San Diego, CA: Academic Press.

McGregor, K. K., Friedman, R. M., Reilly, R. M., & Newman, R. M. (2002). Semantic representation and naming in young children. *Journal of Speech Language and Hearing Research, 45*(2), 332-346.

McKay, A., Davis, C., Savage, G., & Castles, A. (2008). Semantic involvement in reading aloud: Evidence from a nonword training study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1495-1517.

McLennan, C. T. (2007). Challenges facing a complementary-systems approach to abstract and episodic speech perception. *Proceedings of the International Congress of Phonetic Sciences*, 67-70.

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(2), 306-321.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(4), 539-553.

McMurray, B., Spivey, M. J., Aslin, R. N., Tanenhaus, M. K., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology-Human Perception and Performance, 34*(6), 1609-1631.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86*(2), B33-B42.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*, 1113-1126.

Mestres-Misse, A., Camara, E., Rodriguez-Fornells, A., Rotte, M., & Munte, T. F. (2008). Functional Neuroanatomy of Meaning Acquisition from Context. *Journal of Cognitive Neuroscience, 20*(12), 2153-2166.

Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A., & Rosenbaum, R. S. (2006). The cognitive neuroscience of remote episodic, semantic and spatial memory. *Current Opinion in Neurobiology, 16*(2), 179-190.

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., et al. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *Journal of Anatomy, 207*(1), 35-66.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*(4), 379-390.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*(1), 365-378.

Munoz, M., & Insausti, R. (2005). Cortical efferents of the entorhinal cortex and the adjacent parahippocampal region in the monkey (Macaca fascicularis). *European Journal of Neuroscience, 22*(6), 1368-1388.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*(2), 217-227.

Nader, K., & Einarsson, E. O. (2010). Memory reconsolidation: an update *Year in Cognitive Neuroscience 2010* (Vol. 1191, pp. 27-41). Oxford: Blackwell Publishing.

Newman, R. S. (2008). The level of detail in infants' word learning. *Current Directions in Psychological Science, 17*(3), 229-232.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110*(4), 611-646.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*(3), 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204-238.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 3-27.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109-132). San Diego: Academic Press.

Nygaard, L. C., Burt, S. A., & Queen, J. S. (2000). Surface form typicality and asymmetric transfer in episodic memory for spoken words. *Journal of Experimental Psychology-Learning Memory and Cognition, 26*(5), 1228-1244.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech-perception as a talker-contingent process. *Psychological Science, 5*(1), 42-46.

O'Reilly, R. C. (2001). The division of labour between the neocortex and hippocampus. In G. Houghton (Ed.), *Connectionist modeling in cognitive psychology*: Psychology Press.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends in Cognitive Sciences, 6*(12), 505-510.

O'Reilly, R. C., & Rudy, J. W. (2000). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108*(2), 311-345.

Orfanidou, E., Davis, M. H., Ford, M. A., & Marslen Wilson, W. (2011). Perceptual and response components in repetition priming of spoken words and pseudowords. *The Quarterly Journal of Experimental Psychology, 64*(1), 96-121.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology-Learning Memory and Cognition, 19*(2), 309-328.

Penfield, W., & Milner, B. (1958). Memory deficit produced by bilateral lesions in the hippocampal zone. *Archives of Neurology and Psychiatry, 79*(5), 475-497.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137-157). Amsterdam: John Benjamins.

Pilotti, M., Bergman, E. T., Gallo, D. A., Sommers, M., & Roediger, H. L. (2000). Direct comparison of auditory implicit memory tests. *Psychonomic Bulletin & Review, 7*(2), 347-353.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9-32). San Diego: Academic Press.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology-Learning Memory and Cognition, 21*(3), 785-794.

Protopapas, A. (2007). Check Vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behaviour Research Methods, 39*(4), 859-862.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia, 42*(1), 2-13.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology-Human Perception and Performance, 23*(3), 651-666.

Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. P. (2009). Statistical frequency in perception affects children's lexical production. *Cognition, 111*, 372-377.

Riedel, G., Micheau, J., Lam, A. G. M., Roloff, E. V., Martin, S. J., Bridge, H., et al. (1999). Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nature Neuroscience, 2*(10), 898-905.

Roediger, H. L., & Blaxton, T. A. (1987). Effects of varying modality, surface-features, and retention interval on priming in word-fragment completion. *Memory & Cognition, 15*(5), 379-388.

Rosenbaum, R. S., Winocur, G., & Moscovitch, M. (2001). New views on old memories: re-evaluating the role of the hippocampal complex. *Behavioural Brain Research, 127*(1-2), 183-197.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339-349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*(6), 608-635.

Rothkopf, E. Z. (1971). Incidental memory for location of information in text. *Journal of Verbal Language and Verbal Behaviour, 10*, 608-613.

Rueckl, J. G. (1990). Similarity effects in word and pseudoword repetition priming. *Journal of Experimental Psychology-Learning Memory and Cognition, 16*(3), 374-391.

Sadakata, M., & McQueen, J. M. (2011). *The role of variability in non-native perceptual learning of Japanese geminate-singleton fricative contrast.* Paper presented at the Interspeech, Florence, Italy.

Salasoo, A., Shiffrin, R. M., & Feustel, T. C. (1985). Building permanent memory codes: Codification and repetition effects in word identification. *Journal of Experimental Psychology-General, 114*(1), 50-77.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology-Human Perception and Performance, 3*(1), 1-17.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology-Learning Memory and Cognition, 18*(5), 915-930.

Scharenborg, O., Mitterer, H., & McQueen, J. M. (2011). *Perceptual learning of liquids.* Paper presented at the Interspeech, Florence, Italy.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology Neurosurgery and Psychiatry, 20*(1), 11-21.

Sedin, L. M. (2006). *The recognition and acquisition of inflected spoken words: Evidence from phonetic categorization.* Unpublished doctoral dissertation, University of York, York, England.

Sharon, T., Moscovitch, M., & Gilboa, A. (2011). Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America, 108*(3), 1146-1151.

Shatzman, K. B., & McQueen, J. M. (2006). The modulation of lexical competition by segment duration. *Psychonomic Bulletin & Review, 13*(6), 966-971.

Sheffert, S. M. (1998). Contributions of surface and conceptual information to recognition memory. *Perception & Psychophysics, 60*(7), 1141-1152.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 34*(5), 665-685.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*(2), 833-870.

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language, 51*(2), 173-189.

Snodgrass, J. G., hirshman, E., & Fan, J. (1996). The sensory match effect in recognition memory: Perceptual fluency or episodic trace. *Memory & Cognition, 24*(3), 367-383.

Snoeren, N. D., Gaskell, M. G., & Di Betta, A. M. (2009). The Perception of Assimilation in Newly Learned Novel Words. *Journal of Experimental Psychology-Learning Memory and Cognition, 35*(2), 542-549.

Sommers, M. S. (1999). Perceptual specificity and implicit auditory priming in older and younger adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1236-1255.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96*(3), 1314-1324.

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology, 5*(2), 169-177.

Squire, L. R., & Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus, 8*(3), 205-211.

Stickgold, R., & Walker, M. P. (2007). Sleep-dependent memory consolidation and reconsolidation. *Sleep Medicine, 8*(4), 331-343.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech Language and Hearing Research, 44*(6), 1321-1337.

Suzuki, W. A., & Eichenbaum, H. (2000). The neurophysiology of memory. *Parahippocampal Region, 911*, 175-191.

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernandez, G. (2009). Shift from hippocampal to neocortical centered retrieval network with consolidation. *Journal of Neuroscience, 29*(32), 10087-10093.

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M. J., et al. (2006). Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America, 103*(3), 756-761.

Takehara-Nishiuchi, K., & McNaughton, B. L. (2008). Spontaneous Changes of Neocortical Code for Associative Memory During Consolidation. *Science, 322*(5903), 960-963.

Tamminen, J. (2010). *Learning new words: Effects of meaning, memory consolidation, and sleep.* University of York, York.

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.

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.

Teng, E., & Squire, L. R. (1999). Memory for places learned long ago is intact after hippocampal damage. *Nature, 400*(6745), 675-677.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*(3), 339-363.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Oxford University Press.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76*(6), 559-573.

Vanlancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*(2), 195-209.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., VanPaesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277*(5324), 376-380.

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society, 17*, 1-15.

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: context manipulation and hippocampal representations of memory. *Nature Neuroscience, 10*(5), 555-557.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441-517.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *The Journal of Neuroscience, 25*(11), 3002-3008.

Zola-Morgan, S. M., & Squire, L. R. (1990). The primate hippocampal-formation: Evidence for a time-limited role in memory storage. *Science, 250*(4978), 288-290.