The time-course of talker-specificity and lexical competition effects during word learning

Helen Brown* & M. Gareth Gaskell

Department of Psychology, University of York, YO10 5DD, UK

* Please address correspondence to:

Helen Brown

Department of Psychology

University of Warwick

Coventry

CV4 7AL

Email: helen.brown@warwick.ac.uk

Tel: 02476 573 946

Short title: Talker-specificity during word learning

Word count: 10,439 (main text =9266, references = 1173)

Abstract

Three experiments examined the time-course of talker-specificity and lexical competition effects during spoken word learning. Talker-specificity effects depend on access to highly detailed lexical representations, whilst lexical competition may exploit more abstract representations. By tracking the time-courses of these effects concurrently we examined whether there was a common mechanism underlying their storage and retention. Talker-specificity effects on recognition of novel words were robust immediately after study and were generally stable over the course of a week. In contrast, lexical competition effects emerged only at delayed test points. This time-course dissociation supports a dual-system model of lexical processing in which episodic representations of new words are generated rapidly, but robust representations underlying lexical competition emerge only after a period of offline consolidation.

Keywords: word learning, lexical representation, memory consolidation, indexical specificity, lexical competition

Talker-specificity effects (TSEs) occur when spoken words are processed faster and more accurately if heard in the same voice at study and test as opposed to different voices. They suggest that voice-specific details are encoded and stored in memory when a word is heard. Lexical competition, on the other hand, is often characterized as relying on a more abstract phonological code; words that are phonologically similar take part in a competition process during spoken word recognition (Marslen-Wilson & Zwitserlood, 1989). Nonetheless, some models of word recognition encapsulate this competition process using the same kind of episodic or detailed representation that could underlie TSEs (*e.g.,* Goldinger, 1998). In this paper we look at the extent to which the processes that support TSEs and lexical competition have similar properties. In particular, we look at the time-course of the TSEs and lexical competition effects when new words are encountered as a means of determining whether they rely on the same processes for encoding and retention.

With regards to TSEs, many studies have already shown that recently studied existing words are processed faster and more accurately when the surface details of the speech form (Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; Goh, 2005; Goldinger, 1996; Goldinger, Kleider, & Shelley, 1999; McLennan & Luce, 2005; Schacter & Church, 1992; Sheffert, 1998), or the written form (see Tenpenny, 1995, for a review), remain consistent between study and test. Similar specificity effects have also been observed immediately after study for newly-learned words (Creel, Aslin, & Tanenhaus, 2008; Creel & Tumlin, 2009, 2011).

Additionally, studies of lexically-driven perceptual learning (*e.g.,* Norris, McQueen, & Cutler, 2003), in which exposure to an ambiguous phoneme /?/ midway between (for example) /f/ and /s/ in the context of /f/-final words results in a bias to interpret /?/ as /f/ at test (and vice-versa given exposure to /?/ in /s/-final words), also provide further evidence that talker information can be retained in memory. For instance, Eisner and McQueen (2005) found robust perceptual learning only when the same talker was heard during exposure and test. However, Kraljic and Samuel (2005) reported more mixed findings. Perceptual learning was not observed when items were trained in a male voice and tested in a female voice, suggesting that perceptual learning effects were talker specific. Conversely, when items were trained in a female voice and tested in a male voice perceptual learning was observed despite the change in talker, suggesting that perceptual learning was not talker specific (see also Kraljic & Samuel, 2006). Thus, lexically-driven perceptual learning effects appear in some cases to be talker-specific and in others talker-general even when participants are tested immediately after study.

Whereas it is clear from previous research that TSEs can be observed soon after learning there is less evidence relating to whether talker-specific details are retained in long-term memory. The retention intervals used in the experiments cited above were all relatively short (typically less than an hour). Only a handful of studies have examined longer-term retention of episodic details in lexical memory, and these have provided mixed findings. On one hand, Goldinger (1996) found significant TSEs for existing words in an identification-in-noise task one week post-study. Ernestus (2009) also showed that information about unreduced vowels affected recognition of newly-encountered past participles one week post-study. On the other hand, using the same stimuli as in the identification-in-noise task, Goldinger failed to demonstrate sustained TSEs one week post-study in an old/new recognition task, suggesting that episodic details may be lost over time. Thus, whilst there is substantial evidence indicating that talker-specific details are likely to be encoded and affect processing for a short period of time immediately after a word has been encountered, the retention of this information in long-term memory is less clear.

In comparison to the immediacy of TSEs, lexical competition effects for new words are typically absent immediately post-study, emerging only in delayed test sessions (Dumay

& Gaskell, 2007; Henderson, Weighall, Brown, & Gaskell, 2012). One way to examine whether new and existing lexical representations have been integrated is to determine whether new words (*e.g., biscal*) impact upon processing of phonologically-similar existing words (*e.g., biscuit*). Slower processing of existing 'base-words' with new competitors (compared to control 'base-words' without new competitors) indicates that the new nonwords have been integrated with existing knowledge and are engaging in lexical competition with similar-sounding words during spoken word recognition (*cf.* Marslen-Wilson & Zwitserlood, 1989). Previous studies largely demonstrate slowed processing of test base-words only at delayed test points following periods of sleep-associated offline consolidation (Dumay & Gaskell, 2007; Henderson *et al.,* 2012; although see Lindsay & Gaskell, 2013).

To summarise, existing evidence suggests that lexical competition effects for new words are typically absent immediately after learning, and emerge following a consolidation period. On the other hand TSEs are robust immediately after learning, but it is less clear whether they remain in the longer term. The different time-courses observed for TSEs and lexical competition effects might suggest that these two effects depend upon different processing mechanisms. In the context of word learning, a dissociation between immediate establishment of highly-detailed lexical representations and their integration with existing knowledge may be interpreted within dual-system models of learning and memory (*e.g.,* Davis & Gaskell, 2009; McClelland, McNaughton, & O'Reilly, 1995). Such dual-system models assume that new information is initially retained in an episodic temporary store, but over time becomes integrated (via consolidation processes) into a long-term memory store. Assuming that a dual-systems account is viable, there remain uncertainties regarding the precise roles of the two subsystems in the representation and processing of words.

It is possible that the two different subsystems are responsible for two different types of learning: one system for learning specifics and one for learning generalities (O'Reilly &

Norman, 2002). In other words, consolidation may be responsible for generating abstract, context-free representations. Such a finding would be consistent with hybrid models of lexical memory, in which representations are initially episodic, with multiple episodes combining into more abstract units over time, given multiple exposures to a word (Feustel, Shiffrin, & Salasoo, 1983; Goldinger, 2007; Grossberg, 1986; McLennan, Luce, & Charles-Luce, 2005). Importantly, episodic representations are not lost once more abstract representations are formed in hybrid models. Rather, both representations may co-exist in memory. Distributed memory models (*e.g.,* McClelland & Rumelhart, 1985) offer a similar explanation; traces of individual experiences are represented as unique patterns of activation across a number of nodes in a connectionist network, with abstraction emerging from the superposition of similar memory traces.

Alternatively the consolidation processes responsible for integrating new and existing lexical information might strengthen and enhance memory for all encoded information. Evidence that this is the case would support pure episodic/exemplar models of lexical memory which assume that the lexicon consists of episodic traces containing perceptual and contextual details specific to each individual occurrence of a word (Goldinger, 1998; Hintzman, 1986, 1988; Jacoby, 1983a, 1983b).

Although previous research suggests that TSEs and lexical competition effects follow different time-courses, there are a number of factors limiting this conclusion. First, experiments using existing words to examine the retention of talker-specific information are limited by the fact that these words will have been encountered many times, in many different voices, prior to the experimental sessions. TSEs may be masked by this past experience. Second, methodological differences between studies examining TSEs and lexical competition effects make a direct comparison of their time-courses difficult. Studies examining the time-course of TSEs often use a between-participants design in which each participant is tested at only one time point (*e.g.,* Goldinger, 1996) whilst studies examining the time-course of lexical competition effects typically use a within-participants design (*e.g.,* Dumay & Gaskell, 2007). Perhaps crucially, the number of words encountered during study, and the level of exposure to each word tend to be very different. For instance, participants in Goldinger's study encountered 150 existing words just once each before completing a test of TSEs whilst Dumay and Gaskell (2007) exposed participants to 24 new words 36 times each before testing for lexical competition effects. This level of exposure may be crucial in engaging consolidation processes to enhance later memory.

The experiments reported here used an artificial lexicon enabling us to eliminate potential confounds associated with participants having heard existing words in many different voices prior to the experiment. Moreover, by teaching participants new words we were able to carefully control the number of voices that each word was encountered in during study, as well as the number of exposures to each word prior to test. All participants received the same amount of exposure to the same number of words, and completed tests of both TSEs and lexical competition effects at the same time points. This enabled us, for the first time, to test the time-course of TSEs and lexical competition effects in comparable circumstances.

A subsidiary question was whether talker information affected lexical processing once new and existing information had been integrated in long-term memory. Few previous studies have addressed this question. In one experiment Creel *et al.* (2008) taught participants novel words-novel object associations. During study each novel word was heard in only one voice. Critically, the target and competitor items (a novel cohort or rhyme competitor) were spoken either consistently by the same talker, or by different talkers. At test 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 suggest that talker-specific information affected the degree to which two phonologically similar novel words engaged in lexical competition (see Creel & Tumlin, 2009, 2011, for similar findings). Nonetheless, Creel *et al.*'s experiment demonstrates only that talker information affects lexical competition within a small set of novel words; they did not investigate whether talker information affects the amount of lexical competition that is observed between existing and novel words. The design of the three experiments reported here allowed us to address this question.

Returning to the key question, a dual-systems account would predict marked differences in the time-courses of the tasks: episodic TSEs should emerge immediately after study and remain or weaken over time, whereas lexical competition effects should be absent immediately and emerge only after a consolidation period. Alternatively, an episodic/exemplar account would predict that TSEs and lexical competition effects should follow the same time-course, given that they are both underpinned by the same learning process and representation.

Experiment 1

In Experiment 1 participants studied 24 fictitious nonwords (*e.g., biscal*), with half of the items consistently spoken by a male talker and half consistently by a female talker. Talker gender was selected as the episodic detail to be manipulated since it has been widely used in previous studies with existing words. Participants completed two test tasks immediately after study, as well as one day later, and one week later; (i) a lexical decision task, designed to measure lexical competition between new and existing words, and (ii) an old/new categorisation task, used to measure the extent to which talker information affected recognition of the new words themselves. Critically, half of the studied items were heard in different voices at study and test whilst the other half remained in the same voice, enabling us to examine TSEs as well as the extent to which talker identity affected lexical competition between new and existing words.

Method

Participants

Thirty-one students (*age range* 18–23yrs, 9 male) from the University of York completed the experiment, receiving either payment or partial course-credit. Participants in this and subsequent experiments were native British English speakers and reported no known hearing, speech or language impairments. Informed consent was obtained prior to the first session.

Stimuli

Forty-eight stimulus triplets, each containing one existing 'base-word' (*e.g. biscuit*) and two nonwords (*e.g. biscal*, *biscan*), were selected from stimuli used by Tamminen and Gaskell (2008; see Appendix A). Base-words were monomorphemic with uniqueness points located at or before the final vowel. Nonwords differed from their base-word at the final vowel, and from each other at the final consonant/consonant cluster. The nonwords encountered during study will be referred to as '*novel nonwords'* and the untrained nonwords used as distracters in the old/new categorisation task as '*foil nonwords'*. Lexical competition between phonologically similar words occurs up to the point where only one word in the lexicon matches the speech input (the uniqueness point of the word; Cohort Model – Marslen-Wilson & Zwitserlood, 1989). By teaching participants novel nonwords that differ from their base-words only after the uniqueness point of that word, the uniqueness point can be artificially shifted towards the offset of the base-word. Thus, slowed processing of base-words with novel nonword competitors (compared to control base-words without novel competitors) in a lexical decision task would indicate that newly-learned nonwords were engaging in lexical competition.

The stimulus triplets were split into two lists of 24, with base-words matched on initial phoneme, number of syllables (12 bisyllabic and 12 trisyllabic per list), and as closely as possible on number of phonemes (*M*=7.96, *Range*=6-11) and frequency (*M*=3.63, *Range*=2-14; CELEX database, Baayen, Piepenbrock, & van Rijn, 1993). T-tests indicated that whilst the two lists did not differ significantly in the mean number of phonemes per word, t(46)=-.423, p=.67, the difference between the mean frequency of each list approached significance, t(46)=-1.799, p=.08. In order to ensure that this marginal difference did not affect results the two lists were counterbalanced across participants; each list was the ‘test’ list for half of the participants, and the ‘control’ list for the other half.

Forty-eight monomorphemic English nouns (24 monosyllabic, 12 bisyllabic and 12 trisyllabic), 96 nonwords (each generated by changing one or two phonemes of existing words), and 30 practice items (15 words; 15 nonwords) were also selected from Tamminen and Gaskell (2008) as filler items for the lexical decision task.

One male and one female British English speaker recorded the stimuli in a sound attenuated booth. On average, the acoustic duration of words recorded by the male talker (M=691ms, SD=96ms) was shorter than those recorded by the female talker (M=805ms, SD=89ms), t(143)=-13.75, p<.001. Although this difference in acoustic duration between talkers was unplanned, and will have added to the acoustic differences between talkers this is not of critical importance since the experiments reported in this paper primarily address the time-course of TSEs for novel nonwords rather than the specific variables driving the TSEs themselves. The stimuli were digitized at a 44.1Hz sampling rate with 16-bit analogue-to-digital conversion. Peak amplitude was normalised using Adobe Audition.

Design

Participants were tested individually in sound-attenuated booths. Tasks were run using DMDX experimental software (Forster & Forster, 2003). Stimuli were presented binaurally over headphones at a comfortable listening level.

Participant completed three sessions on Days 1, 2 (~24 hours later), and 8 (one week later). In Session 1 participants were exposed to the novel nonwords in a phoneme monitoring task. Participants subsequently completed the lexical decision and old/new categorisation tasks. On Days 2 and 8 participants completed only the lexical decision and the old/new categorisation tasks. The order of these two tasks was fixed across sessions, with the lexical decision task always occurring first.

Procedure

Each participant studied one list of 24 novel nonwords, counterbalanced across participants, in a *phoneme monitoring task*. Twelve nonwords were spoken consistently by the male talker and 12 by the female talker. Participants listened for and indicated the presence/absence of specified phonemes in the novel nonwords. Following five existing-word practice trials, participants completed six experimental blocks, each specifying a different target phoneme (/p/, /t/, /b/, /m/, /s/, /d/). The novel nonwords occurred three times per block, with the order of the novel nonwords randomised in groups of 24 (*i.e.,* one full repetition of the list).

Unless otherwise stated, in this and all subsequent tasks, (i) instructions emphasised both speed and accuracy; (ii) feedback stating the average RT and the number of errors made was provided at the end of each block to encourage quick and accurate responding; (iii) 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.

At test the study lists were further subdivided so that 6 of the 12 nonwords studied in the male voice were tested in the female voice, whilst the other 6 remained in the male voice, and likewise for nonwords studied in the female voice. Overall, 12 nonwords were heard in the same voice as study, and 12 in a different voice. The test voice was the same for all items within a stimulus triplet; if the novel nonword ‘*biscal*’ was spoken in a male voice at test, then the foil nonword '*biscan*' was also heard in the male voice in the old/new categorization task, and the base-word '*biscuit*' was spoken in the male voice in the lexical decision task. Test-talker remained constant across test-sessions such that items classed as different-talker items in Session 1 remained in the opposite voice to study at all test-points (and likewise for all same-talker items). Participants were not informed about the manipulation of voices between study and test in order to avoid drawing attention to this variable.

In the *lexical decision task* participants heard all 48 base-words, 48 word fillers, and 96 nonword fillers presented in a randomised order in two experimental blocks of 96 items. Blocks were matched in the number of test base-words, control base-words, word fillers, and nonword fillers, with the order of the blocks counterbalanced across participants. Half of the test base-words, control base-words, word fillers, and non-words were heard in the male voice, and half in the female voice, counterbalanced across participants so that half heard each item in the male voice and half in the female voice. Thirty practice items were included at the start of 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 response key respectively.

In the *old/new categorisation task* participants heard the 24 studied novel nonwords and 24 corresponding foil nonwords presented one at a time in a randomised order. Participants decided whether each nonword was old (heard during the phoneme monitoring task) or new (had never been heard before), indicating their response by pressing the left or right response key respectively. Feedback was not provided in this task.

Results

RTs were analysed for the lexical competition task, whilst accuracy data were analysed for the old-new categorisation task. For all analyses in this and 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.

Study phase

Fifteen participants learned List 1 and 16 learned List 2. The mean error rate in the phoneme monitoring task was 5.6% (*SD*=2.5%).[1] A 2 (*study talker*: male, female) x 2 (*list*: 1, 2) repeated-measures ANOVA revealed a non-significant main effect of list, Fs<1, but a marginally significant effect of study talker, $F_1(1,28)=3.51$, p=.071, $\eta_p^2=.11$, $F_2(1,46)=1.83$, p=.18; $\eta_p^2=.04$, indicating that participants made fewer errors in the phoneme monitoring task when items were heard in the male (*M*=5.0%) rather than the female (*M*=6.0%) voice. Nonetheless, there was no significant interaction between study talker and list, Fs<1.

Lexical competition effects

Across all items in the lexical decision task the mean accuracy score was 92.3% (*SD*=4.0%), indicating that participants were paying close attention to the task. Only data from the 48 base-words were included in the analysis, allowing comparison between words with (*test*) and without (*control*) novel nonword competitors. All incorrect responses were removed prior to analysis (6.5% of the data), as were correct data points with RTs <200ms or >2.5*SD* from the mean RT for each participant in each session (2.3% of the data). One participant had an error score more than 2.5*SD* above the mean and was removed from the dataset. With this participant removed the mean RT was 930ms (*SD*=214ms) and the mean accuracy was 91.7%. Mean RTs for test and control base-words in each session are reported in Table 1. Difference scores are plotted in Figure 1a.

A 2 (*base-word type:* test, control) x 3 (*day:* 1, 2, 8) repeated-measures ANOVA revealed a significant main effect of day, $F_1(2,56)=13.47$, p<.001, $\eta_p^2=.33$, $F_2(2,94)=45.82$, p<.001, $\eta_p^2=.49$. RTs were significantly slower on Day 1 (*M*=968ms) compared to both Day 2 (*M*=904ms), $F_1(1,28)=34.87$, p<.001, $\eta_p^2=.56$, $F_2(1,47)=109.34$, p<.001, $\eta_p^2=.70$, and Day

8 (*M*=927ms), $F_1(1,28)=7.00$, p=.013, $\eta_p^2=.20$, $F_2(1,47)=32.52$, p<.001, $\eta_p^2=.41$, likely due either to practice effects and task repetition resulting in faster RTs on Days 2 and 8, or to fatigue on Day 1 where participants had just completed the 20-25 min phoneme monitoring task. RTs were also significantly faster on Day 2 compared to Day 8, $F_1(1,28)=4.95$, p=.035, $\eta_p^2=.15$, $F_2(1,47)=10.64$, p=.002, $\eta_p^2=.19$.

The main effect of word-type was not significant, $F_1(1,28)=1.08$, p=.31, $\eta_p^2=.04$, $F_2(1,47)=1.30$, p=.26, $\eta_p^2=.03$, but crucially the interaction between day and word-type was highly significant, $F_1(2,56)=10.64$, p<.001, $\eta_p^2=.28$, $F_2(2,94)=5.46$, p=.006, $\eta_p^2=.10$. RTs were quicker to test compared with control base-words on Day 1, $F_1(1,28)=6.49$, p=.017. $\eta_p^2=.19$, although this effect was only marginally significant by-items, $F_2(1,47)=3.16$, p=.082, $\eta_p^2=.06$. In comparison, test word RTs were significantly slower than control word RTs on Day 2, $F_1(1,28)=8.28$, p=.008, $\eta_p^2=.23$, $F_2(1,47)=7.47$, p=.009, $\eta_p^2=.14$, with this effect remaining marginally significant by-participants on Day 8, $F_1(1,28)=3.31$, p=.079, $\eta_p^2=.11$, although the effect was no longer significant by-items, $F_2(1,47)=2.04$, p=.16, $\eta_p^2=.04$, suggesting that lexical competition between phonologically similar nonwords and base-words emerged after a period of sleep-associated offline consolidation and remained (to some degree) one week later. The facilitatory effects observed for test base-words on Day 1 (see also Gaskell & Dumay, 2003) may be due to participants having become aware of the phonological similarity between the studied novel nonwords and their base-words. Even if participants were not consciously aware of this similarity, hearing the novel nonwords 18 times during study may have primed their phonologically similar base-words such that they were activated more rapidly than control base-words on Day 1.

Whilst the analyses above show that, as predicted, lexical competition effects emerged only after a consolidation period, they do not distinguish between base-words heard in the same or a different voice to the studied nonwords. If consolidation preserves information about the talker then stronger lexical competition effects should be evident in cases where base-words and new competitors were matched both in terms of the initial phoneme sequence *and* the identity of the talker. To address this possibility, control-test difference scores were calculated separately for same- and different-talker base-words (see Table 2). These difference scores were analysed using a 2 (*base-word talker:* same, different) x 3 (*day:* 1,2,8) repeated-measures ANOVA.. As in the main analysis there was a significant main effect of day, $F_1(2,56)=9.97$, $p<.001$, $\eta_p^2=.26$, $F_2(2,94)=4.52$, $p=.013$, $\eta_p^2=.09$. The main effect of base-word talker was not significant, $F_1<1$, $F_2(1,47)=1.30$, $p=.26$, $\eta_p^2=.03$. There was however a significant interaction between day and base-word talker in the by-participants analysis, $F_1(2,56)=3.27$, $p=.045$, $\eta_p^2=.10$, although this was not significant by items, $F_2(2,94)=1.83$, $p=.17$, $\eta_p^2=.04$. Further analysis revealed that the difference between same and different talker items approached significance only on Day 8, $F_1(1,28)=3.07$, $p=.09$, $\eta_p^2=.10$, $F_2(1,47)=5.85$, $p=.020$, $\eta_p^2=.11$ (see Figure 1b).

Talker-specificity effects

In the old/new categorisation task, participants responded correctly to 83.7% (*SD*=7.0%) of the items. The data were analysed using signal detection theory (*SDT*; Green & Swets, 1966): d-prime (d′) provides an estimate of sensitivity (the ability to distinguish signal from noise) that is unaffected by individual response biases (Figure 2). Hit rates and false alarm rates are reported in Table 3.

A 2 (*test-talker*: same, different) x 3 (*day*: 1, 2, 8) repeated-measures ANOVA revealed a significant main effect of test-phase talker, $F(1,29)=36.46$, $p<.001$, $\eta_p^2=.56$, with higher accuracy scores for same- compared to different-talker items. There was also a marginal main effect of day, $F(2,58)=3.05$, $p=.055$, $\eta_p^2=.10$. Posthoc comparisons revealed a significant difference only between Days 2 and 8, $F(1,29)=4.50$, $p=.043$, $\eta_p^2=.13$. All other between-session comparisons were non-significant. Critically, there was no interaction between test-phase talker and day, F(2,58)=1.60, p=.21, η_p^2=.05, suggesting that the same-talker advantage did not change significantly over time.

Discussion

Experiment 1's key finding was that TSEs and lexical competition effects for novel words have different time-courses. Talker identity influenced recognition of the new words immediately after study, and remained equally influential at delayed test-points up to one week later. Conversely, lexical competition effects were absent immediately after learning, but emerged after a period of offline consolidation and remained (to some degree) one week later. This dissociation may be explained by these two effects being subserved by different processes/memory mechanisms. Talker identity can be encoded immediately into new episodic lexical representations, whereas lexical competition effects may be reliant on lexical representations that arise as a result of the consolidation processes responsible for integrating new and existing information.

The effect of time on the emergence of lexical competition is consistent with previous research showing that a period of sleep-associated offline consolidation is sufficient for phonologically-similar new and existing words to become integrated and begin competing during spoken word recognition (Davis, Di Betta, Macdonald, & Gaskell, 2009; Dumay & Gaskell, 2007; Dumay, Gaskell, & Feng, 2004; Gaskell & Dumay, 2003; Henderson *et al.*, 2012; Tamminen & Gaskell, 2008). Experiment 1 also provided an opportunity to address whether lexical competition effects for novel words were talker-specific. There was limited evidence that this was the case. The magnitude of lexical competition effects did not differ significantly for same- and different-talker items on Day 2, and even on Day 8 this difference only approached significance by-participants despite being fully significant by-items, suggesting that competition between similar-sounding novel and existing words during spoken word recognition may rely primarily on more abstract phonological information. This finding is inconsistent with results from Creel *et al.*'s (2008) study in which talker-specific lexical competition effects were observed between pairs of phonologically similar novel words. However, in this study participants were required only to consider a small set of newly learned words associated with visually presented novel objects. In such an environment talker information may become more salient or even strategically encoded and used to aid performance at test. Our experiment, which examined lexical competition between existing and novel words, required participants to consider the whole lexicon, making strategic use of talker information less likely.

In terms of TSEs in recognition memory, Experiment 1 showed that accurate form-based representations of new words were rapidly generated, with both phonological and talker information being encoded and stored. The immediacy of TSEs is consistent with previous studies demonstrating TSEs for existing words immediately after study (*e.g.,* Goldinger, 1996). However, the finding that TSEs remained strong in the d′ data one week later is more novel, contrasting with Goldinger's (1996) old/new categorisation task in which TSEs for existing words declined over the course of a week. The retention of TSEs over a week in Experiment 1 is particularly surprising given that test-talker remained constant across test sessions. Hearing a novel nonword in a different voice at test on Day 1 should have resulted in that nonword being represented by two unique memory traces, each containing different talker information. As a result, recognition of these different-talker items should have subsequently improved on Days 2 and 8, thus decreasing the size of the TSEs. The absence of an interaction between test-talker and day suggests that TSEs did not change in size over a week.

One explanation for the different time-courses of TSEs for existing and novel words may be that repeated presentation of a single token of each nonword during study highlighted idiosyncrasies in these tokens, encouraging deliberate/strategic encoding of talker identity.

An additional experiment conducted in our lab using a similar methodology provides initial evidence against this suggestion. This experiment used a surprise old/new categorisation task, reducing the likelihood that participants deliberately encoded talker information during study as a cue to aid later recognition memory. Moreover, participants completed test sessions only on Days 1 and 8, minimizing re-testing effects and potential confounds associated with using a within-participants design. Nevertheless, d′ data from the old/new categorization task still revealed significant TSEs in the Day 8 re-test. These results suggest that talker information is automatically encoded and stored alongside phonological information when new words are encountered.

An alternative explanation highlights the use of nonwords in our study. Only one token of each nonword was heard during study in Experiment 1 and participants had no prior experience of these items. In contrast, participants would have encountered different tokens of each existing word many times prior to Goldinger's study. As more tokens of the same word are encountered, the invariant properties of the word may become abstracted and episodic details lost; that is, variability in the input may be essential (in addition to periods of sleep-associated offline consolidation) in order for robust abstract lexical representations to be established in long-term memory. Experiments 2 and 3 explore this possibility in more detail.

There is already evidence from other research areas within psycholinguistics that variability is vital in order for robust abstract representations that are capable of generalisation to be formed. For instance, variable input during training appears to be vital in order for adults to form robust, abstract perceptual categories (*e.g.,* Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999), and developmental studies suggest that variability in the input may be important in the early stages of word learning (Singh, Morgan & White, 2004; Rost & McMurray, 2009). Thus, it seems plausible that increasing the variability of study tokens may alter the time-course with which abstract representations become dominant in our word learning experiment. That is, introducing greater talker-variability during study may alter both the time-course of TSEs and the lexical competition effects. However, different types of variability during study may have different effects on these two measures.

If talker information remains constant across different training instances of a novel word while speech rate and intonation differ (*within-talker variability*; Experiment 2) then TSEs may still be observed when participants are later required to recognise the studied novel nonwords. Within an exemplar model of lexical memory this should arise due to all episodic traces of each novel nonword containing the same talker information. Within a hybrid model this pattern of TSEs may result from the variable aspects of the input (speech rate and intonation) being treated as irrelevant, but the invariant aspects of the input (in this case talker identity, in addition to the phonological form of the novel word) being retained in memory. This results in a further (but rather extreme) prediction that TSEs in recognition memory may become stronger over time as a 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. If this latter prediction is correct then talker-specific lexical competition effects may also be observed in Experiment 2.

In comparison, introducing multiple talkers during study (*between-talker variability;* Experiment 3) should result in smaller TSEs than Experiment 1. Within an exemplar memory model this would result from different episodic traces of a novel word containing different talker information. An exemplar model would predict that the size of TSEs following between-talker variability should be equivalent in size at all time points since recognition memory and lexical competition should be based on activation of the same set of traces at all test points. Alternatively, a hybrid model of lexical representation might predict that if talker identity varies across training tokens alongside variation in speech rate and intonation 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. If abstract representations require a period of offline consolidation in order to become established, then the size of TSEs may change between the immediate test, and delayed test sessions in Experiment 3.

Experiment 2: Within-talker variability

Method

Participants

Thirty-two undergraduate students (*age range* 18-21yrs, 18 male) from the University of York completed the experiment.

Stimuli

The stimuli were the same as Experiment 1. Eighteen tokens of each novel nonword (varying in intonation and articulation rate) were recorded by each talker for the phoneme monitoring task. 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, base-words, and filler items for the lexical decision task were also re-recorded to minimise differences in recording or voice quality. As in Experiment 1 the acoustic duration of items recorded by the male talker (M=694ms, SD=158ms) was, on average, shorter than those recorded by the female talker (M=830ms, SD=151ms), t(1151)=-5.76, p<.001. However, the standard deviations reported here indicate that there was a similar amount of variability in the duration of tokens produced by both talkers. Audio stimuli were recorded and edited in the same manner as in Experiment 1. The counterbalancing of stimuli and talker was identical to Experiment 1 except for the lexical decision task (changes to this task are described below).

Design and procedure

The *phoneme monitoring* task was identical to Experiment 1, with half of the words consistently spoken by the male talker and half consistently by the female talker. The key difference was that 18 different tokens of each novel nonword (all produced by the same talker) were heard rather than a single token being repeated 18 times. Training tokens were ordered by acoustic duration, and split into three groups of six tokens (slow, medium, and fast). Within each of the six phoneme monitoring blocks one slow, one medium, and one fast token of each nonword was heard. The order of these three tokens was randomised within each block.

As in Experiment 1, participants completed the lexical decision and old/new categorisation tasks on Days 1, 2, and 8. The lexical decision task was identical to Experiment 1 except that half of the participants heard only the female talker and half heard only the male talker. For all participants this manipulation still resulted in half of the base-words being heard in the same voice as the corresponding studied novel nonword, and half being heard in a different voice, enabling talker-specific lexical competition effects to be examined. The old/new categorisation task was identical to Experiment 1.

Results

Study Phase

Sixteen participants learned each list. Two items were removed from the data set due to a programming error that resulted in these items having greater than/less than 18 exposures during the study task. With these items removed the mean error rate was 5.1% (*SD*=2.3%). A 2 (*study talker*: male, female) x 2 (*list*: 1, 2) repeated-measures ANOVA 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 slightly more errors 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, Fs<1, nor was there a significant interaction between list and study talker, $F_1<1$, $F_2(1,44)=1.64$, p=.21, η_p^2=.04. As such, whilst participants made more errors to List 2 items, study talker did not influence this. List is therefore unlikely to have influenced TSEs in the test tasks.

Lexical competition effects

Overall accuracy in the lexical decision task was 92.6% (*SD*=5.8%), indicating that participants were paying close attention to the task. As in Experiment 1, only RT data from the 48 base-words were analysed. Data points were removed if they corresponded to incorrect responses (6.7% of the dataset), if they contained data points with RTs <200ms or >2.5*SD* from the mean of each participant in each session (2.5% of the data). One participant had an error score more than 2.5*SD* above the grand mean and was removed from the dataset. With this participant removed the mean RT was 987ms (*SD*=195ms) and accuracy was 91.5%.

A 2 (*base-word type:* test, control) x 3 (*day:* 1, 2, 8) repeated-measures ANOVA for the RT data (Table 1 and Figure 3) revealed a significant main effect of day, F_1(2,58)=6.94, p=.002, η_p^2=.19, F_2(2,90)=35.88, p<.001, η_p^2=.44, with RTs being significantly slower on Day 1 (*M*=1019ms) compared to Days 2 (*M*=976ms) and 8 (*M*=979ms), as in Experiment 1. The main effect of base-word type was not significant, F_1(1,29)=1.24, p=.28, η_p^2=.04, F_2<1, nor was the interaction between base-word type and day, F_1(2,58)=2.21, p=.12, η_p^2=.07, F_2(2,90)=1.52, p=.23, η_p^2=.03. Follow up comparisons confirmed that there were no significant differences between test and control base-words on Days 1 and 8, Fs<1. However, there was a significant main effect of base-word type on Day 2, F_1(1,29)=5.61, p=.025, η_p^2=.16, F_2(1,45)=4.11 p=.049, η_p^2=.08, with slower RTs to test than control base-words at this time point as expected. However, since the lexical competition effects were not fully significant in the overall analysis the data were not further subdivided to look at the talker-specificity of lexical competition.

Talker-specificity effects

In the *old/new categorisation* task the mean accuracy score was 84.8% (*SD*=7.2%). Analysis of d′ data (Figure 4, Table 3) using a 2 (*test-phase talker:* same, different) x 3 (*day*: 1, 2, 8) repeated-measures ANOVA revealed a significant main effect of test-phase, F(1,30)=12.66, p=.001, η_p^2=.30, but a non-significant main effect of day, F(2,60)=1.41, p=.25, η_p^2=.05, and no interaction between test-phase talker and day, F<1, indicating that the size of TSEs in recognition memory did not change over time.

Discussion

Experiment 2, like Experiment 1, demonstrated that detailed representations of novel nonwords were rapidly generated and could support significant TSEs in recognition memory up to one week post-exposure. Thus, introducing within-talker variability did not change the time-course of TSEs for new words.

In terms of lexical competition, robust effects were present on Day 2 consistent with previous research showing that new words engage in competition within similar sounding words only after a period of offline consolidation (e.g., Dumay & Gaskell, 2007; Henderson et al., 2012). However, lexical competition effects were absent on Day 8, and the interaction between day and base-word type was not statistically significant in the main ANOVA, contrary to expectations. This finding is interesting given that all previous studies examining the engagement of newly learned words in lexical competition with similar-sounding existing words have used a single repeated token during the study phase of the experiment. It is possible that variability in the training tokens weakens the extent to which new words engage in lexical competition over the longer term. This possibility will be discussed further below.

Experiment 3: Between-talker variability

Method

Participants

Thirty-two undergraduate students (*age range* 18-23yrs, 10 male) from the University of York completed the experiment.

Stimuli

The stimuli consisted of the 48 word triplets from Experiment 1. For the phoneme monitoring task nine tokens of each novel nonword were selected from the tokens recorded by the two talkers (M1, F2) from earlier experiments. Two additional speakers (1 male (M2), 1 female (F1)), also recorded nine tokens of each novel nonword, varying in intonation and articulation rate. Stimuli were recorded and edited as described above. The mean acoustic duration of novel nonword tokens produced by each talker are as follows: Male 1 – M=705ms, SD=173ms; Male 2 – M=629ms, SD=144ms; Female 1 – M=825ms, SD=175ms; Female 2 – M=832ms, SD=162ms. The test-phase stimuli were identical to those used in Experiment 2.

Design and procedure

In the phoneme monitoring task each novel nonword was heard in two voices (1 male, 1 female). In order that the test materials from Experiment 2 could be used, the male speaker from Experiments 1/2 (M1) was paired with the new female speaker (F1), and the female speaker from Experiments 1/2 (F2) was paired with the new male speaker (M2). Thus, in the phoneme monitoring task half of the novel nonwords were spoken consistently by M1/F1, and half consistently by M2/F2. Items were encountered 18 times during study, with 9 tokens spoken by each talker. As in Experiment 2, tokens from each talker were ordered by acoustic duration, and were split into three groups (slow, medium, and fast). In each block of phoneme monitoring one slow, one medium, and one fast token of each novel nonword was heard, presented in a random order. All four speakers were included in each block 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 (*e.g.,* if the first block contained 2 female tokens and 1 male token, the second block contained 2 male tokens and 1 female token *etc.*). Counterbalancing of talkers and stimuli was the same as in Experiments 1 and 2 except that on occasions when participants had previously only heard M1 during study they now heard tokens from both M1 and F1. Likewise, on occasions when participants had previously heard only F2 during study they now heard tokens from both F2 and M2. The counterbalancing of test stimuli was identical to Experiment 2, involving only talkers M1 and F2.

The test phase of the experiment was identical to Experiment 2. As before, half of the items were heard in the same voice as study (*e.g.,* items studied in M1/F1 were heard in voice M1 at test) and half were heard in a different voice (*e.g.,* items studied in voices M1/F1 were heard in voice F2 at test).

Results

Study Phase

Sixteen participants learned each list. The mean phoneme monitoring error rate was 5.4% (*SD*=2.1%). A 2 (*study talker*: M1/F1, M2/F2) x 2 (*list*: 1, 2) repeated-measures ANOVA revealed non-significant main effects of study-phase talker, Fs<1, and list, $F_1(1,30)=1.69$, p=.20, $\eta_p^2=.05$, $F_2(1,46)=1.37$, p=.25, $\eta_p^2=.03$, as well as a non-significant interaction between study-phase talker and list, $F_1(1,30)=2.88$, p=.10, $\eta_p^2=.09$, $F_2(1,46)=1.20$, p=.28, $\eta_p^2=.03$.

Lexical competition effects

Overall participants responded correctly to 91.3% (*SD*=4.7%) of items in the lexical decision task. Data from the 48 base-words were filtered as in Experiment 1; incorrect responses (8.2%) and data points with RTs <200ms or >2.5SD above the mean RT for each participant in each session (2.3%) were removed prior to analysis. One participant with an error score more than 2.5*SD* above the grand mean was removed from the dataset. With this participant removed the mean RT was 1005ms (*SD*=247ms) and the accuracy was 90.0%. Mean RTs for test and control base-words in each session are reported in Table 1 and difference scores are plotted in Figure 5.

A 2 (*base-word type*: test, control) x 3 (*day*: 1, 2, 8) repeated-measures ANOVA revealed a significant main effect of day, $F_1(2,58)=3.69$, $p=.031$, $\eta_p^2=.11$, $F_2(2,94)=13.62$, $p<.001$, $\eta_p^2=.23$. RTs were significantly slower on Day 1 (*M*=1024ms) compared to Day 2 (*M*=983ms), $F_1(1,29)=7.93$, $p=.009$, $\eta_p^2=.22$, $F_2(1,47)=28.41$, $p<.001$, $\eta_p^2=.38$. All other comparisons were non-significant. The main effect of base-word type was not significant by participants $F_1(1,29)=1.47$, $p=.24$, $\eta_p^2=.05$, although it did approach significance by items, $F_2(1,47)=3.63$, $p=.063$, $\eta_p^2=.07$. The interaction between base-word type and day was not significant, $F_1(2,58)=1.18$, $p=.32$, $\eta_p^2=.04$, $F_2<1$. Separate analyses of the data from each test session confirmed these findings, revealing non-significant main effects of base-word type at all time points. Thus, although numerically there were some hints of competition effects on Days 2 (11ms) and 8 (16ms), neither of these reached significance level. As in Experiment 2, since evidence of lexical competition was not reliable in the overall analyses, the data were not further subdivided in order to examine talker-specificity in the lexical competition measures.

Talker-specificity effects

In the *old/new categorisation* task participants responded correctly to 83.8% (*SD*=7.4%) of the items. 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 2 (*test-phase talker*: same, different) x 3 (*day*: 1, 2, 8) repeated-measures ANOVA revealed a significant main effect of test-talker for d′ values (Figure 6, Table 3), $F(1,28)=7.55$, $p=.01$, $\eta_p^2=.21$. The main effect of day was also significant, $F(2,56)=3.47$, $p=.038$, $\eta_p^2=.11$, with a d′ scores being significantly higher on Day 1 compared to Day 8, F(1,28)=5.14, p=.031, η_p^2=.16 and on Day 2 compared to Day 8, F(1,28)=4.46, p=.044, η_p^2=.14. There was no difference between d′ scores on Days 1 and 2, F<1. There was also no interaction between test-phase talker and day, F(2,56)=1.23, p=.30, η_p^2=.04, suggesting that the size of TSEs in recognition memory did not change significantly over time.

Discussion

Experiment 3 demonstrates that even when novel nonwords are heard in more than one voice during study, robust representations containing both phonological and talker information are still formed and are able to support significant TSEs in recognition memory up to one week post-exposure. Nonetheless, overall old/new categorisation performance did decrease significantly over time, unlike Experiment 2 (within-talker variability) and Experiment 1 (no variability; marginally significant decrease, p=.055). This finding suggests that the representations supporting old/new categorisation decisions may decay at a faster rate after novel words have been heard in more than one voice compared to when they are studied in a single voice.

The presence of significant TSEs in Experiment 3 in the absence of a significant interaction between test-talker and day is important for two reasons. Firstly, it argues against Geiselman & Crawley's (1983) *voice connotation* hypothesis, which states that TSEs should only be observed when two talkers of different genders are used (see Palmeri, Goldinger, & Pisoni, 1993, for further evidence that TSEs can be observed even when participants encounter multiple male and multiple female voices). In Experiment 3, TSEs must depend upon retention of talker-specific details, not simply the presence of different gender tags associated with each novel nonword. Secondly, only nine tokens of each novel nonword were heard in each study voice in Experiment 3 (compared to 18 in Experiments 1 and 2). Thus, if

Experiment 3 had revealed non-significant TSEs it could have been argued that these arose due to less robust representation of talker information in memory. This was not the case.

Lexical competition data in Experiment 3 revealed numerical trends towards lexical competition on Days 2 and 8, but these effects did not reach statistical significance. There are now many published studies reporting significant delayed lexical competition effects, although as would be expected from the "dance of the p-values" there are also a smaller number of studies that have not reached the somewhat arbitrary $p<.05$ cut-off for the delayed competition effect (Cumming, 2014). It is possible that the relatively weak evidence for lexical competition in Experiments 2 and 3, as compared with Experiment 1 is just another example of equivalent underlying effects happening to land either side of the significance cut-off. Nonetheless it is also possible that the weakness of lexical competition effects across Experiments 2 and 3 may in this case stem from the increased variability between training tokens; when multiple talkers are heard the variability in voice information may attract attention that would otherwise be used for other cognitive processes (Martin, Mullennix, Pisoni, & Summers, 1989). In the case of learning new words, talker variability during study may make the task of generating robust phonological representations of the new words more difficult. This explanation could also account for the decreased accuracy in the old/new categorisation task over the course of a week post-study in Experiment 3.

Given the lack of statistically reliable lexical competition effects within Experiments 2 and 3 it was not possible to explore whether talker information affected the magnitude of lexical competition effects. Nonetheless in both experiments there were non-significant numerical trends in the data that were consistent with weak lexical competition effects on Days 2 and 8. Below we present a combined analysis of lexical decision data from Experiments 1, 2, and 3, which provides a more powerful assessment of the time-course of lexical competition effects as well as determining whether these effects are affected by talker identity.

Cross-Experiment Analysis: Lexical Competition

RT data for the 48 base-words in Experiments 1, 2, and 3 were combined, and were analysed in a 2 (*base-word type*: test, control) x 3 (*day*: 1, 2, 8) x 3 (*variability*: none, within-talker, between-talker) repeated-measures ANOVA to determine the robustness of overall lexical competition effects (Table 1 and Figure 7a). There were significant main effects of day, $F_1(2,172)=20.34$, $p<.001$, $\eta_p^2=.19$, $F_2(2,90)=69.41$, $p<.001$, $\eta_p^2=.61$ (with all pairwise comparisons revealing significant differences – Day 1=1002ms; Day 2=953ms; Day 8=970ms), and variability, $F_1(2,86)=3.23$, $p=.044$, $\eta_p^2=.07$, $F_2(2,90)=83.80$, $p<.001$, $\eta_p^2=.65$, (reflecting differences in the overall mean RTs for each experiment – Experiment 1 = 930ms; Experiment 2 = 987ms; Experiment 3 = 1001ms). The main effect of base-word type was also marginally significant, $F_1(1,86)=3.71$, $p=.057$, $\eta_p^2=.04$, $F_2(1,45)=3.24$, $p=.079$, $\eta_p^2=.07$.

Most importantly, the critical interaction between day and base-word type was significant, $F_1(2,172)=7.92$, $p=.001$, $\eta_p^2=.08$, $F_2(2,90)=4.53$, $p=.013$, $\eta_p^2=.09$. Further analysis revealed that there was no significant difference between RTs to test and control base-words on Day 1, $F_1(1,86)=1.34$, $p=.25$, $\eta_p^2=.02$, $F_2<1$, but that this difference was significant on Day 2, $F_1(1,86)=12.37$, $p=.001$, $\eta_p^2=.13$, $F_2(1,45)=11.72$, $p=.001$, $\eta_p^2=.21$, and marginally significant (by participants only) on Day 8, $F_1(1,86)=3.65$, $p=.06$, $\eta_p^2=.04$, $F_2(1,45)=2.13$, $p=.15$, $\eta_p^2=.05$, 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, 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).

Following on from these findings a second set of analyses examined whether there was any evidence of talker-specific lexical competition effects (Table 2 and Figure 7b). As in Experiment 1, test-control difference scores were calculated separately for same- and different-talker base-words. These difference scores were then analysed in a 2 (*base-word talker:* same, different) x 3 (*day:* 1, 2, 8) x 3 (*variability*: none, within-talker, between-talker) repeated-measures ANOVA. There was once again a main effect of day, $F_1(2,172)=7.57$, $p=.001$, $\eta_p^2=.08$, $F_2(2,90)=4.20$, $p=.018$, $\eta_p^2=.09$. However, the main effects of variability, Fs<1, and base-word talker $F_1(1,86)=1.03$, $p=.31$, $\eta_p^2=.01$, $F_2<1$ were both non-significant. None of the interactions reached or approached significance in both by-participants and by-items analyses, although some reached significance in one or the other. Together these findings do not provide good evidence for talker information being preserved in the representations underlying lexical competition. Nonetheless, there are some numerical trends in the data (see Table 2) that are suggestive of an influence of talker-specific information affecting lexical competition effects. Further investigation is required in order to fully rule out the possibility that talker information affects lexical competition between newly-learned words and phonologically similar existing words.

General Discussion

The key finding in these experiments is that TSEs and lexical competition effects for newly learned words follow different time-courses and as such may rely on different processing and/or memory mechanisms. TSEs for novel words were present immediately after exposure and remained stable during the week post-study. There was no evidence of any consolidation benefit (*i.e.,* strengthening of TSEs at later test points). In contrast, evidence that newly-learned words engaged in lexical competition with phonologically similar existing words was absent immediately after the new words had been learned, but emerged following a period of sleep-associated offline consolidation, as evidenced by the interactions between day and base-word type in Experiment 1 and the combined analysis. The independent time-courses of TSEs and lexical competition effects provide compelling evidence that they are underpinned by separate mechanisms. This result is inconsistent with a purely episodic model of the mental lexicon (*e.g.,* Goldinger, 1998). Instead, the results are compatible with hybrid models of lexical memory which assume that two different representational systems (episodic and abstract) co-exist in memory.

A complementary learning systems framework (McClelland *et al.*, 1995) offers one account of a hybrid model, in which isolated and integrated representations depend upon two different subsystems. Evidence that TSEs affect recognition of newly learned words immediately after study suggests that isolated representations of new words in the first subsystem can be generated rapidly and are detailed in nature, maintaining talker information in addition to phonological information. In contrast, the lexical competition data (from Experiment 1 and the cross-experiment analysis) suggest that more extended periods of offline consolidation are required in order for new representations to become robustly integrated with existing knowledge in the second subsystem. The absence of strong evidence supporting talker-specific effects on lexical competition measures, although a null effect, provides some evidence that the subsystem underlying lexical competition effects may rely on more abstract representations than those involved in the simple recognition of new words.

This stands in contrast with research by Creel *et al.* (2008), which demonstrated significant talker-specific lexical competition between pairs of newly learned phonologically-similar nonwords, as well as between pairs of recently encountered phonologically-similar existing words. A critical difference between Creel *et al.*'s study and the experiments reported here is that we exposed participants only to the novel nonwords, not their phonologically-similar base-words, during study. It may be that stronger talker-specific lexical competition effects would emerge if both the novel nonwords and their existing base-words were encountered in the same voice during the study phase of the experiment. If this were the case it would suggest that talker specific lexical competition effects are dependent upon the presence of both detailed isolated episodic representations (that are available only if words have been recently encountered) in addition to robust abstract representations in long-term lexical memory.

Notably TSEs during recognition of the new words did not appear to either strengthen or weaken over the course of a week, indicating that the detailed, isolated representations are maintained for at least one week after a new word has been encountered, even once new lexical knowledge has been integrated into long term lexical memory. This finding is consistent with hybrid memory models in which episodic and abstract representations are able to co-exist, but inconsistent with previous research suggesting that TSEs for existing words decrease over time in a similar old/new categorisation task (*e.g.,* Goldinger, 1996). One explanation for this difference may be that repeating the nonwords 18 times in the same voice/pair of voices during study in the current experiments may have strengthened memory for talker information in comparison to Goldinger's study in which participants encountered each existing word only once before test. Notably a study by Ernestus (2009), in which participants were exposed to 12 tokens of each items during familiarisation, also demonstrated retention of detailed lexical representations one week post study. Alternatively the difference between Goldinger's study and the experiments reported here may be that existing words will have been encountered many times, in many different voices prior to an experiment, and that this experience may subsequently mask or weaken TSEs for these items.

To summarise, the current data show a clear dissociation in the time-course of the emergence of two key aspects of lexical knowledge. New words appear to be initially encoded in a form that retains detailed episodic information such as talker identity. Representations in this episodic subsystem can be maintained for at least a week after new words are initially learned, but do not show any consolidation advantage when tested at later time points. In contrast, engagement in lexical competition is absent immediately after learning, but emerges following a consolidation period of 24 hours. This profile of learning suggests that lexical competition is dependent on the consolidation of rapidly formed (episodic) representations into a more integrated network that links similar sounding new and existing words. Notably, the consolidation of new lexical representations into the integrated network did not appear to trigger the decay of episodic representations underlying recognition memory. Rather, the two types of representation appear to be able to co-exist, supporting the independent time-course of TSEs and lexical competition effects during word learning and providing support for a hybrid or 'dual-system' model of lexical memory.

Acknowledgements

We would like to thank Alex Reid and Lisa Henderson for their help in producing the audio recordings for Experiment 3. The research was supported by an Economic and Social Research Council (ESRC) studentship awarded to HB, and an ESRC Grant awarded to MGG (ES/E015263/1). Old/new categorisation data from these experiments were presented in the following 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 Interspeech 2011, ISCA,* Florence, Italy.

References

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

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. doi: 10.3758/BF03206911

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. doi: 10.3758/BF03206883

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.

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. doi: 10.1016/j.cognition.2007.03.013

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. doi: 10.1016/j.jml.2011.06.005

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. doi: 10.1177/0956797613504966

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*, *21*(4), 803-820. doi: 10.1162/jocn.2009.21059

Davis, M.H., & Gaskell, M.G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London, Series B - Biological Sciences, 364,* 3605-3800. doi: 10.1098/rstb.2009.0111

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

Dumay, N., Gaskell, M.G., & Feng, X. (2004). *A day in the life of a spoken word*. Paper presented at the Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society., Mahwah, NJ.

Eisner, F., & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224-238. doi: 10.3758/BF03206487

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. doi: 10.1037/0096-3445.112.3.309

Forster, J.C., & Forster, K.I. (2003). DMDX: A Windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, & Computers, 35*, 116-124. doi: 10.3758/BF03195503

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

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. doi: 10.1037/0278-7393.31.1.40

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. doi: 10.1037/0278-7393.22.5.1166

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279. doi: 10.1037/0033-295X.105.2.251

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., 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. doi: 10.3758/BF03211416

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

Grossberg, S.D. (1986). The adaptive self-organization of serial order in behaviour: Speech, language and motor control. In E. C. Schwab, & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines. Speech erpception* (Vol.1, pp.187-294). New York: Academic Press. doi: 10.1016/B978-0-12-631403-8.50011-4

Henderson, L.M., Weighall, A., Brown, H., & Gaskell, M.G. (2012). Vocabulary acquisition is associated with sleep in children. *Developmental Science*, *15,* 674-687. doi: 10.1111/j.1467-7687.2012.01172.x

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428. doi: 10.1037/0033-295X.93.4.411

Hintzman, D.L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528-551. doi: 10.1037/0033-295X.95.4.528

Jacoby, L.L. (1983a). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning Memory and Cognition, 9*(1), 21-38. doi: 10.1037/0278-7393.9.1.21

Jacoby, L.L. (1983b). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior, 22*(5), 485-508. doi: 10.1016/S0022-5371(83)90301-8

Kraljic, T., & Sameul, A.G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141-178. doi: 10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268. doi: 10.3758/BF03193841

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

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accesing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576-585. doi: 10.1037/0096-1523.15.3.576

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. doi: 10.1037/0278-7393.15.4.676

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. doi: 10.1037/0033-295X.102.3.419

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. doi: 10.1037/0096-3445.114.2.159

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. doi: 10.1037/0278-7393.31.2.306

McLennan, C.T., Luce, P.A., & Charles-Luce, J. (2005). Representation of lexical form: Evidence from studies of sublexical ambiguity. *Journal of Experimental Pscyhology: Human Perception and Performance, 31*(6), 1308-1314. doi: 10.1037/0096-1523.31.6.1308

Norris, D., McQueen, J.M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204-238. doi: 10.1016/S0010-0285(03)00006-9

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. doi: 10.1016/S1364-6613(02)02005-3

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. doi: 10.1037/0278-7393.19.2.309

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. doi: 10.1037/0278-7393.21.3.785

Rost, G.C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339-349. doi: 10.1111/j.1467-7687.2008.00786.x

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. doi: 10.1037/0278-7393.18.5.915

Sheffert, S.M. (1998). Contributions of surface and conceptual information to recognition memory. *Perception & Psychophysics, 60*(7), 1141-1152. doi: 10.3758/BF03206164

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. doi: 10.1016/j.jml.2004.04.004

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. doi: 10.1080/17470210701634545

Tenpenny, P.L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*(3), 339-363. doi: 10.3758/BF03210972

Appendix A: Base-words, novel nonwords, and foil nonwords used in Experiments 1-3

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

Footnotes

[1] Error and RT data from the phoneme monitoring task were lost for one participant in Experiment 1 due to a technical fault that occurred at the end of the task. However, since the participant had completed the phoneme monitoring task prior to the fault data from this participant were still included in the lexical decision and old/new categorisation analyses.

Table 1: Mean RTs (in ms) to test (with a novel nonword competitor) and control (without a novel nonword competitor) words in the lexical decision task.

	Day 1		Day 2		Day 8	
	Test	Control	Test	Control	Test	Control
Exp 1	954	974	911	889	933	919
Exp 2	1015	1014	982	962	976	977
Exp 3	1026	1028	990	979	1015	999
Cross-Exp Analysis	999	1006	962	944	975	966

Table 2: Mean RTs (in ms) to same-talker test words, different-talker test words, and control words in the lexical decision task.

	Day 1			Day 2			Day 8		
	Same	Different	Control	Same	Different	Control	Same	Different	Control
Exp 1	958	951	974	904	918	889	949	918	919
Exp 2	1011	1019	1014	997	966	962	980	973	977
Exp 3	1030	1021	1028	998	982	979	1010	1020	999
Cross-Exp Analysis	1000	998	1006	967	956	944	980	971	966

Table 3: Mean hit rates and false alarm rates to same- and different-talker items in the old/new categorisation task.

		Hit Rate			False Alarm Rate		
		Day 1	Day 2	Day 8	Day 1	Day 2	Day 8
Exp 1	Same	.92	.91	.88	.17	.17	.19
	Different	.71	.76	.72	.14	.13	.16
Exp 2	Same	.90	.91	.87	.17	.17	.16
	Different	.81	.81	.77	.12	.12	.14
Exp 3	Same	.90	.85	.85	.18	.18	.17
	Different	.83	.83	.79	.19	.17	.23

Figure 1: (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) base-words in the lexical decision task (Experiment 1). (b) Lexical decision data split according to whether the test base-word was spoken in either the same voice that the corresponding novel word was trained in, or a different voice. Values below 0 indicate the presence of increased lexical competition for test base-words. Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

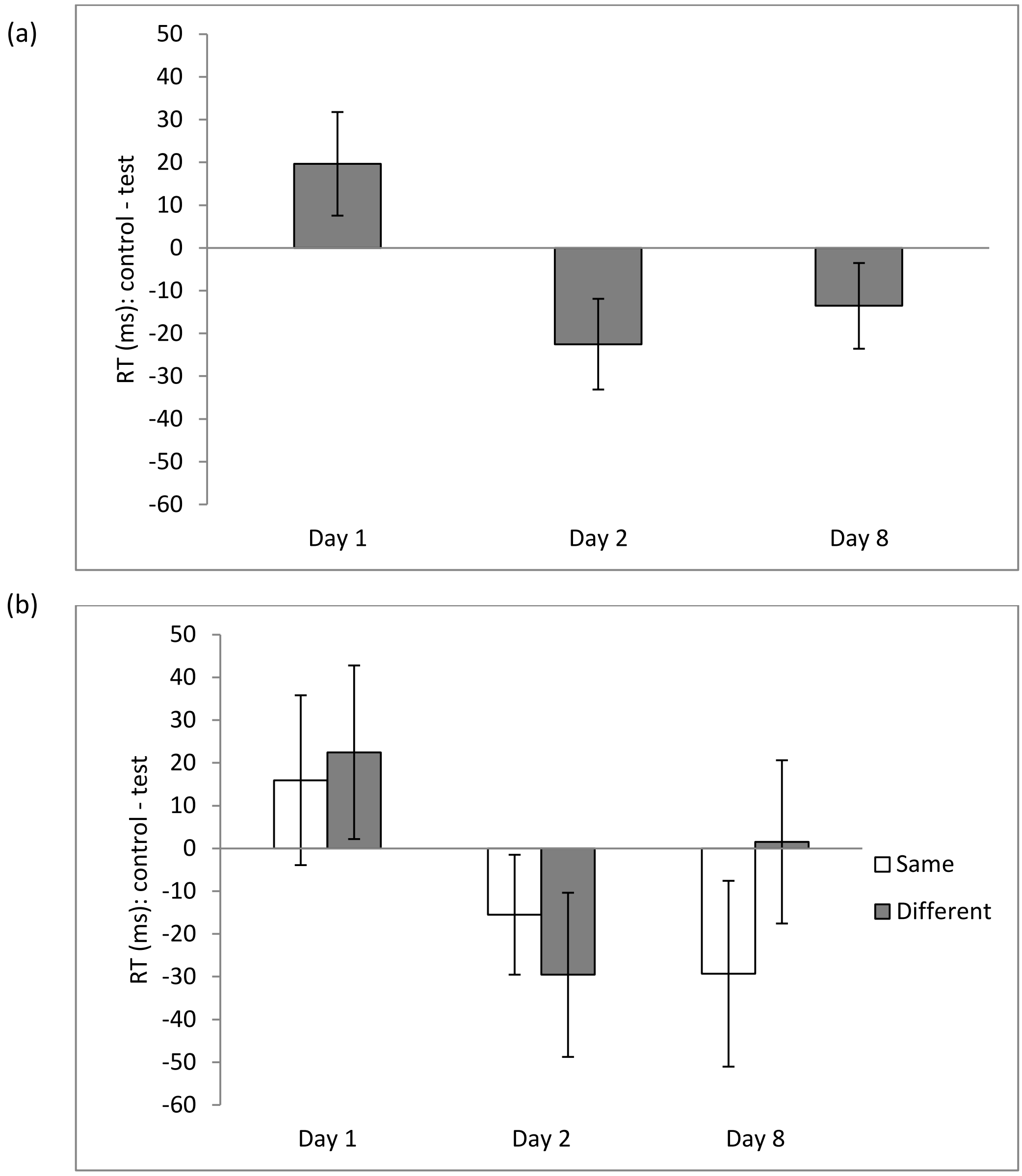


Figure 2: Sensitivity in the old/new categorisation task as a function of whether the study and test talkers were the same or different (Experiment 1). Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

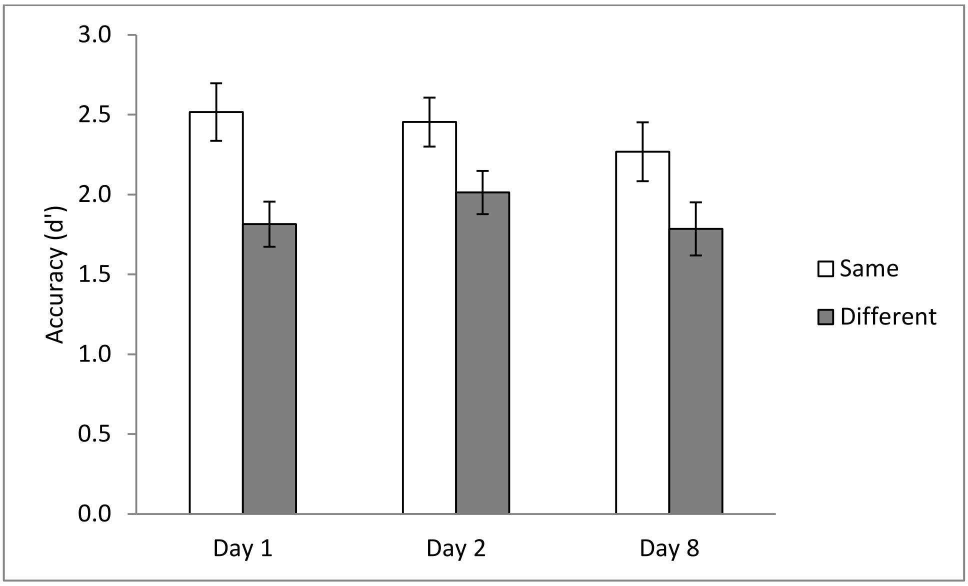


Figure 3: Mean difference between RTs to control (no novel competitor) and test (novel competitor) base-words in the lexical decision task (Experiment 2). Values below 0 indicate the presence of increased lexical competition for test base-words. Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

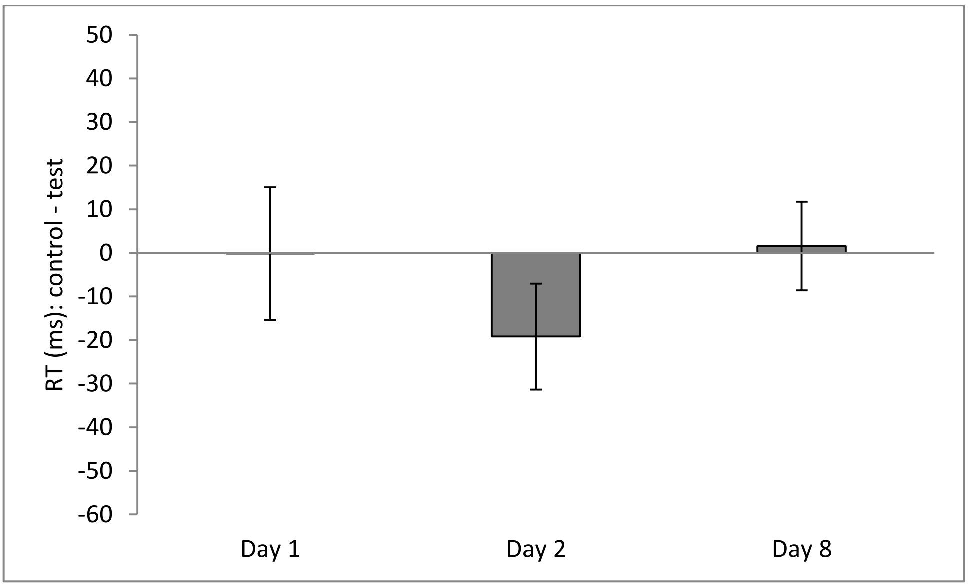


Figure 4: Sensitivity in the old/new categorisation task as a function of whether the study and test talkers were the same or different (Experiment 2). Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

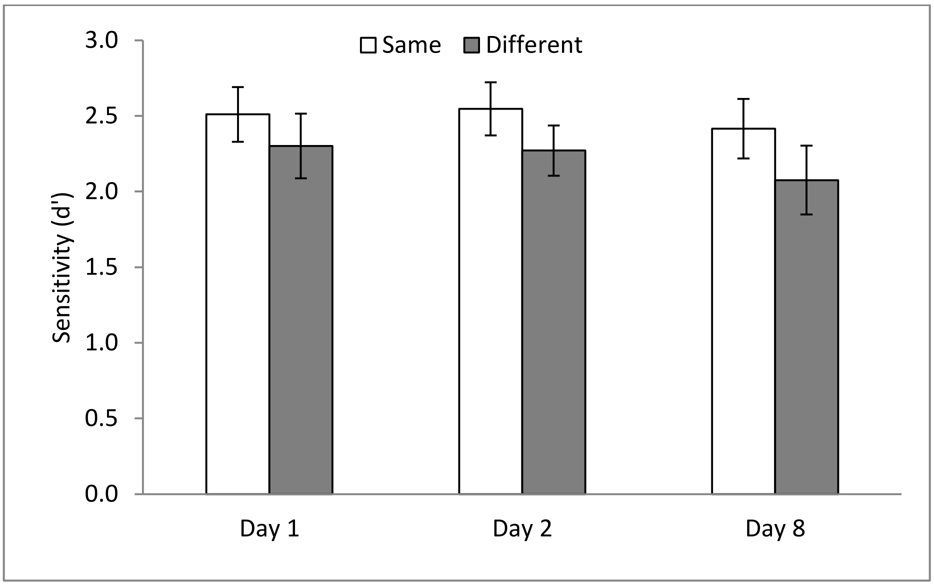


Figure 5: Mean difference between RTs to control (no novel competitor) and test (novel competitor) base-words in the lexical decision task (Experiment 3). Values below 0 indicate the presence of increased lexical competition for test base-words. Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

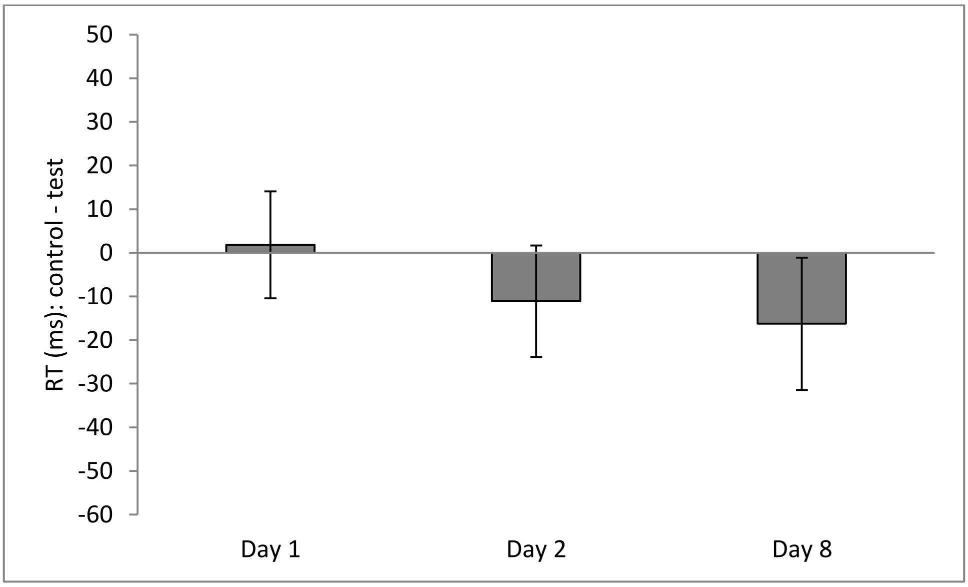


Figure 6: Sensitivity in the old/new categorisation task as a function of whether the study and test talkers were the same or different (Experiment 3). Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).

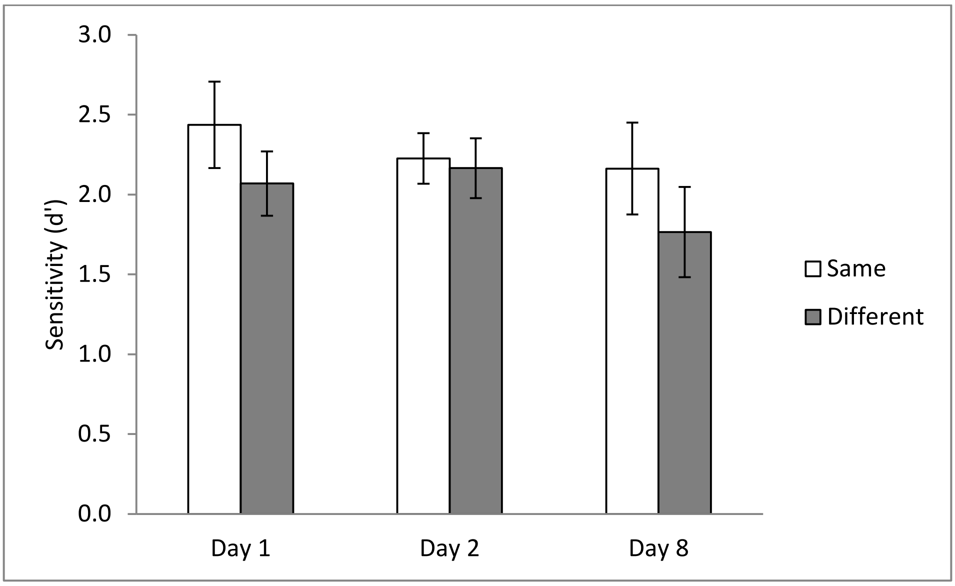
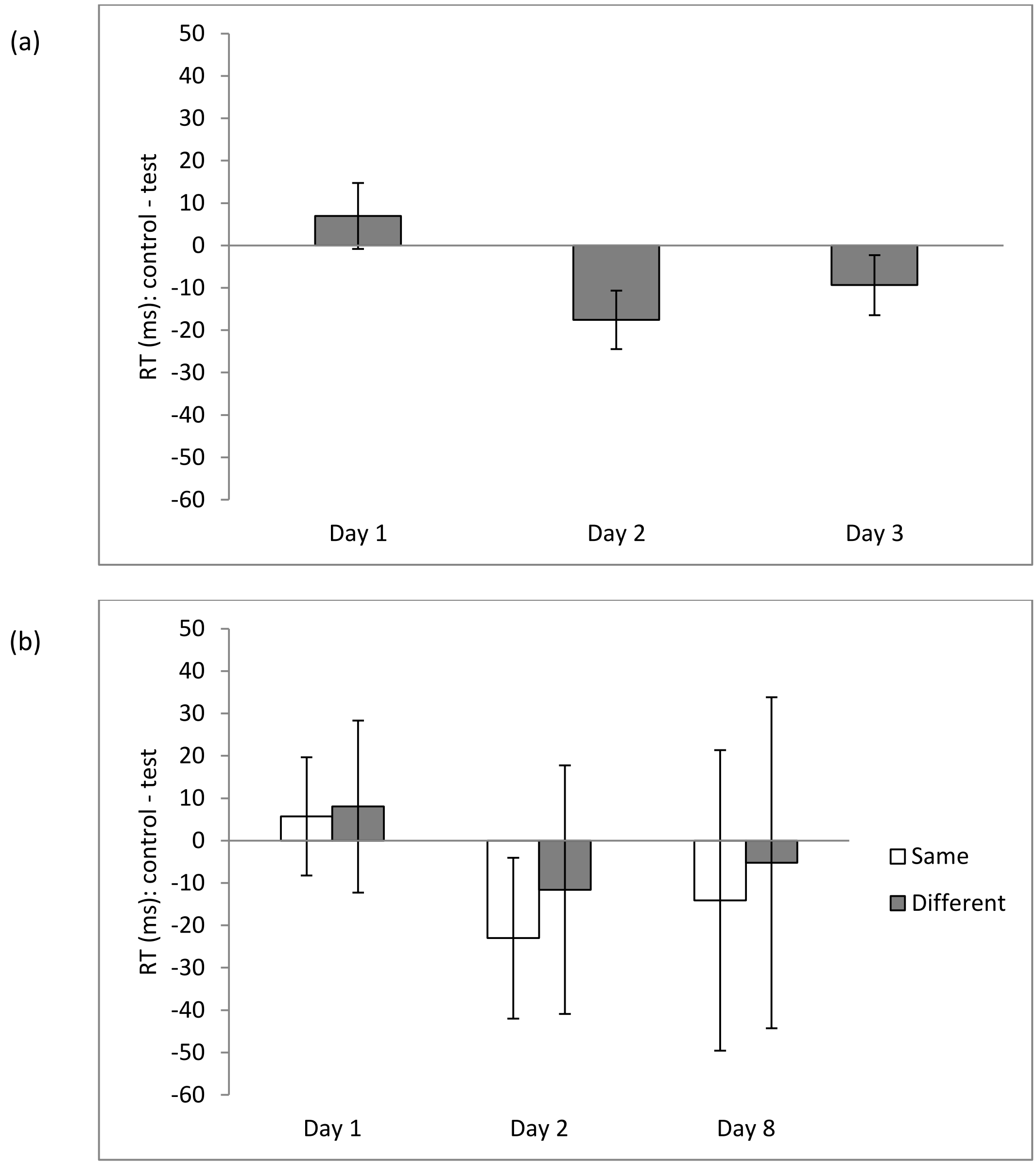


Figure 7: (a) Mean difference between RTs to control (no novel competitor) and test (novel competitor) base-words in the lexical decision task (Experiments 1, 2, and 3 combined). (b) Lexical decision data split according to whether the test base-word was spoken in either the same voice that the corresponding novel word was trained in, or a different voice. Values below 0 indicate the presence of increased lexical competition for test base-words. Error bars indicate 95% confidence intervals after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Cousineau, 2007).


(a)
RT (ms): control - test
Day 1
Day 2
Day 3
(b)
RT (ms): control - test
Day 1
Day 2
Day 8
□ Same
▨ Different