UNIVERSITY OF HULL

DEPARTMENT OF PSYCHOLOGY

The nature of novel word representations

computer mouse tracking shows evidence of immediate lexical engagement effects in adults

being a thesis submitted for the degree of

Doctor of Philosophy

in

Psychology

at the

University of Hull

Submitted by

Mr Andrew Philip Lucas, BSc

Date of first submission: 23rd April 2021

Viva voce passed: 2nd July 2021

Corrections submitted: 29th September 2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

Like any doctoral work, producing this thesis has been difficult, especially due to the CoViD-19 pandemic. However, it has been an immense pleasure and privilege to study for a doctorate, and not something I would have previously thought possible for 'someone like me'. I offer my thanks and gratitude to the following:

- My primary supervisor, Professor Kevin Riggs, for his nurturing and supportive approach to supervision and academia, and for recognising that completing a PhD is so much more than writing a thesis. Also to my secondary supervisors, Drs Richard O'Connor and Shane Lindsay, for their support and patience, and for always being available to discuss my work;
- To the participants in my studies, again, for offering their time, thus allowing science to shuffle forward;
- To researchers who have been kind enough to share their insights and materials, particularly Drs Anna Weighall and Efthymia Kapnoula;
- To the academic and technical staff of the psychology department at the University of Hull;
- To the University of Hull, for providing me with funding, extended during the CoViD-19 pandemic, without which, I could never have afforded postgraduate study;
- To the Sir James Reckitt's charity for their generous grant allowing me to pay participants.

Special thanks should also go to my girlfriend, Şefika Özer, for putting up with my fretting when things were not going well, for celebrating my successes when they were, and for providing me the time and space to blather on and think aloud whenever I needed to.

Lastly, I thank my grandfather, Colin Lucas, for his counsel, for giving me my love of science, and for emphasising the value of education.

THE NATURE OF NOVEL WORD REPRESENTATIONS

COMPUTER MOUSE TRACKING SHOWS EVIDENCE OF IMMEDIATE LEXICAL ENGAGEMENT EFFECTS IN ADULTS

Andrew Philip Lucas

Abstract

Simplistically, words are the mental bundling of a form and a referent. However, words also dynamically interact with one another in the cognitive system, and have other so-called 'lexical properties'. For example, the word 'dog' will cue recognition of 'dock' by shared phonology, and 'cat', by shared semantics. Researchers have suggested that such *lexical engagement* between words emerges slowly, and with sleep. However, newer research suggests that this is not the case. Herein, seven experiments investigate this claim.

Fast mapping (FM), a developmental word learning procedure, has been reported to promote lexical engagement before sleep in adults. Experiment 1 altered the task parameters and failed to replicate this finding. Experiment 2 attempted a methodological replication – again, no effect was found. It is concluded that the effect reported is not easily replicable.

Other findings of pre-sleep lexical engagement were then considered using a novel methodology – computer mouse tracking. Experiments 3 and 4 developed optimal mouse tracking procedures and protocols for studying lexical engagement. Experiment 5 then applied this methodology to novel word learning, and found clear evidence of immediate lexical engagement. Experiment 6 provided evidence that participants were binding the word form to the referent in these pre-sleep lexical representations. Experiment 7 sought to strengthen this finding, but has been postponed due to the CoViD-19 pandemic.

The results are discussed in the context of the distributed cohort model of speech perception, a complementary learning systems account of word learning, and differing abstractionist and episodic accounts of the lexicon. It is concluded that the results may be most clearly explained by an episodic lexicon, although there is a need to develop hybrid models, factoring in consolidation and abstraction for the efficient storage of representations in the long term.

Visualisation of the most common words in this thesis

Part I

Word learning: interfacing language and memory

CHAPTER

ONE

INTRODUCTION

What is *language*, and how do humans acquire it? This is an important question, as language is one of only a few uniquely human capacities (Hockett, 1960; Pinker, 1995; Rivas, 2005). The subject has also been the focus of intense scholarly interest, with over four and a half million Google Scholar results for the keyword 'language', as of September 2021. Language's fundamental function is to impart concepts from one mind to another (Pinker, 1995); however, how humans acquire this ability has been a matter of debate (e.g., Chomsky, 1959; Skinner, 1957). In oral language, a central problem of language acquisition is the learning of words, particularly as it occurs at such a prodigious rate. For example, English-speaking humans learn at least a thousand words yearly for the first twenty years of life (Nation & Waring, 1997). Words are central to oral language, as they label concepts. Thus, it follows that words are an excellent place to begin to understand language and linguistic phenomena. Human word learning is the topic of this thesis.

1.1 What is a word?

As the 'building blocks' of language, a word can be defined as the mental bundling of a concept and a label. The label, properly called a *form*, varies with communication modality. Theoretically, concepts are represented as all the knowledge an individual has relating to that word (Gaskell & Marslen-Wilson, 1997; McClelland, McNaughton & O'Reilly, 1995). An exemplar of a concept referred to by a particular form is that form's 'referent'.

For example, the word 'cat' may be modelled as the phonological code, /kæt/, tied to the collective representation CAT[1] (theorised to be averaged and abstracted across many experienced instances; McClelland et al., 1995). Alone, however, the bundled representation of form and concept is not useful: isolated words a language do not make. Instead, words must be cognitively interconnected, in order to facilitate their productive use. The dynamic cognitive interaction of words is one of their

[1]Note that the use of small capitals font will be used throughout this thesis consistently where a concept is being referred to; forms will either be in IPA transcription where necessary, or else in the standard font.

behavioural signatures (Gaskell & Marslen-Wilson, 1997; Leach & Samuel, 2007; Magnuson, Tanenhaus, Aslin & Dahan, 2003), and distinguishes them from nonsense syllables, or other sounds (which may still carry semantic meaning, e.g., a dog's bark; Bartolotti et al., 2020). It is a question of particular interest how such 'word-like' properties emerge, as they present the most stringent test of a newly-learnt form's lexical status (Davis & Gaskell, 2009; Gaskell & Dumay, 2003; Leach & Samuel, 2007; Lindsay & Gaskell, 2010; McMurray, Kapnoula & Gaskell, 2017).

1.1.1 A framework for studying word learning: 'word-like' properties

Leach and Samuel (2007) made the case that, prior to their work, relatively few authors had distinguished clearly between word learning insofar as it dealt acquiring declarative knowledge (e.g., that a cat is a mammal; that the word 'cat' is spelt with one 'a'), and those dynamic and interactive properties that distinguish words from other sounds. They therefore introduced a new theoretical framework, distinguishing 'lexical configuration' (knowledge about a word's form and meaning) from 'lexical engagement' (e.g., the capacity of a word to facilitate/inhibit recognition of other items in the lexicon). This framework is useful for its theoretical neutrality, and Leach and Samuel's framework allows for a more 'data-driven' approach, the need for which has been recently emphasised (e.g., Cooper, Greve & Henson, 2019c, 2019d; Kapnoula & Samuel, 2019). In the context of this thesis, the 'lexical configuration' and 'lexical engagement' vocabulary will be used throughout.

Regardless of theoretical accounts detailing exactly *how* word learning proceeds, the following must take place. First, a word must 'get in' to the cognitive system, during **encoding**. An important part of this first step is identifying the new word as novel, and forming an internal representation of the to-be-learnt information (i.e., of the novel form and/or referent). Encoding can be well accounted for by models of speech perception, of which there are several (e.g., Gaskell & Marslen-Wilson, 1997; Goldinger, 1998, 2007; Hickok & Poeppel, 2007; Luce & Pisoni, 1998; Norris, 1994; McClelland & Elman, 1986). Secondly, the neuroscientific, behavioural and computational accounts of word learning converge on further processing, during **storage** (Davis & Gaskell, 2009; Kumaran, Hassabis & McClelland, 2016; Goldinger, 2007; Lindsay & Gaskell, 2010; McClelland et al., 1995; McClelland, 2013; McClelland, McNaughton & Lampinen, 2020; McMurray et al., 2017; Palma & Titone, 2020). This remains true even in papers which show that further processing may be a 'sufficient but not necessary' requirement for the emergence of lexical properties (e.g., Fernandes, Kolinsky & Ventura, 2009; Kapnoula & Samuel, 2019; Lindsay & Gaskell, 2013; Weighall, Henderson, Barr, Cairney & Gaskell, 2017). Lastly, a word must be **retrieved** from storage in order to be used productively. A participant's ability to retrieve a newly-learnt word is assessed with different psychological tasks, each of which have particular properties, produce different patterns of data, and assess different aspects of lexical representation.

1.2 Structure of the present thesis

Part I begins with Chapter 2 (p. 7), setting out word learning according to the above framework. Models for encoding and recognising speech are discussed, in addition to a model explaining how words may be stored. Lastly, there is a reflection on the way that retrieval is measured, describing the operationalisation of the lexical configuration and engagement framework (Leach & Samuel, 2007). Chapter 3 (p. 21) concludes Part I by reviewing the recent word learning literature, excluding the most recent developments, and the literature around child and adult fast mapping. Chapter 3 aims to contextualise and provide evidential support for the models discussed in Chapter 2, rounding off this part of the thesis.

Part II introduces a word learning paradigm that has been the subject of some recent interest: 'fast mapping' (FM; e.g., Cooper et al., 2019c). A review of the recent FM literature in children and adults was undertaken, and is presented in Chapter 4 (p. 41). Particularly interesting was the suggestion in the FM literature that usual storage processes (cf., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995) do not occur under FM conditions (e.g., Coutanche & Thompson-Schill, 2014; Sharon, Moscovitch & Gilboa, 2011). The implication of this is that novel words may become 'word-like' faster. The first study of the thesis (Experiment 1, Chapter 5, p. 55; and Experiment 2, Chapter 6, p. 65) attempted to extend and replicate these findings, but were unable to do so.

The course of thesis was therefore changed for Part III. First, a review of the literature considering similar effects outside of FM was performed (Chapter 7, p. 83). The next chapter (Chapter 8, p. 93) discusses mouse tracking; a suitable methodology for further studies in the same field. However, as there was no expertise in running mouse tracking studies locally, two pilot studies needed to be run in order to develop experimental protocols and data analysis skills (Study 2; Experiments 3 and 4; Chapters 9 and 10, pp. 107 and 129).

Mouse tracking was then applied successfully to novel word learning in the final part of the thesis, Part IV (Study 3). This study comprises of Experiments 5–7. Chapter 11 (p. 21) sets out the relationship between these three experiments, and how they came about; Experiments 5–7 themselves are then presented (Chapters 12 to 14; pp. 155, 177 and 191). The thesis is then closed by Chapter 15 (p. 195), which contains the general discussion.

1.3 The impact of the CoViD-19 pandemic

The Doctoral College has advised a short statement on the impact of the CoViD-19 pandemic be included for all affected projects. At the first lockdown, Experiments 1–6 had been completed, with Experiment 7 still to run. Thesis submission was delayed over the summer and through the winter of 2020, with a view to running the experiment. Unfortunately, research activity did not resume, and in January 2021, the university made the decision that projects should be re-configured, not extended further. The thesis is hereby submitted as-is, with the intention to complete, and publish, Experiment 7 as soon as restrictions allow.

CHAPTER

TWO

MODELS OF SPEECH AND MEMORY

2.1 Encoding a novel word: how are words recognised?

Encoding is the process by which external stimuli are represented internally. It is for this reason that models of speech perception, which detail how speech is recognised, best formalise the encoding phase of word learning. A key aspect of encoding for word learning is the recognition of novelty.

2.1.1 Models of speech perception

Whilst there are several models of speech perception in the literature (for a review, see Hickok & Poeppel, 2007), this thesis will focus on the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997). There are three main reasons for this:

1. It makes use of distributed representations, accessed in parallel (for discussion, see Rogers & McClelland, 2014), which are a better conceptual fit for lexical representations (see Gow & Olson, 2015);

2. It explicitly models and recognises the semantic nature of words (Gow & Olson, 2015);

3. It has found support in literature reviews and models of word learning (Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Lindsay & Gaskell, 2010).

Each of these will be discussed in detail below, ahead of a summary of the DCM.

Distributed representations (1)

Computational models use either localised or distributed representations. From a localist perspective, a word is recognised when its node neural network is activated (e.g., McClelland & Elman, 1986). Other words/nodes in the network may need to be inhibited to allow for a target word/node to become activated. By contrast, distributed representations are represented in the network by a *pattern* of many activated nodes (e.g., Gaskell & Marslen-Wilson, 1997). Each of these nodes may represent some component of the word, but no single unit represents the word itself.

Localist accounts create a potential problem insofar as they necessarily redefine a 'word'. As discussed under the heading below, there is recent evidence demonstrating that the activation of semantic information is integral to word recognition, and certain models of speech perception do not account for these newer data (e.g., McClelland & Elman, 1986). Single units cannot capture this multi-dimensional nature of words.

An additional advantage of distributed representations is that unlike models which require an explicit structuring and sorting of information (e.g., McClelland & Elman, 1986), the DCM is not so theoretically constrained. Sub-patterns within the larger representation may be activated by direct mappings from low-level (i.e., perceptual) information, meaning that processing does not need to take place in discrete stages (Gaskell & Marslen-Wilson, 1997).

Semantics in speech perception (2)

Whilst *how* the DCM represents semantics will be discussed later, that it does model semantic processing as part of the word recognition process is an advantage over other models (Spivey, 2016). Gow and Olson (2015) demonstrated that semantics do act in word recognition.

In English phonology, voicing (vibration of the vocal cords) is an important feature in categorising phonemes (compare voiced /d/ to its unvoiced equivalent /t/). To discover if a particular phonemic contrast exists in a language, one looks for so called 'minimal pairs', whereby the changing of a single phoneme in a word changes its meaning. One such minimal pair is formed by 'dusk' (/dʌsk/) and 'tusk' (/tʌsk/): the positions of the articulators to sound these two words are identical, and /dʌsk/ is distinguished from /tʌsk/ only by voicing.

With the creation of an ambiguous consonant /$\frac{\text{d}}{\text{t}}$/ between voiced /d/ and unvoiced /t/, Gow and Olson (2015) provided participants with a form which could not be interpreted on the basis of phonology alone. Instead, participants could only recognise the word through semantics imparted by a sentential context. Participants were played pairs of sentences, and told to press a button if a target sound (e.g., /d/) was present. Example sentences were "The moon rises just at /$\frac{\text{d}}{\text{t}}$ʌsk/", and "A walrus was missing a /$\frac{\text{d}}{\text{t}}$ʌsk/". In response to these examples, 90% of the time participants categorised the ambiguous form /$\frac{\text{d}}{\text{t}}$ʌsk/ in a way congruent with the semantics of sentence – for example, interpreting it as 'dusk', and detecting /d/, in the first sentence, but not in the second.

This suggested that participants were recruiting semantics in order to recognise a word ('dusk' or 'tusk'). However, Gow and Olson (2015) anticipated a challenge to this account, and wondered if participants' responding was truly driven by semantics acting *in* word recognition. An alternative account would state that following the failure of (perceptual) word recognition processes to return a value (absence or presence of the target phoneme /d/), a higher-order control process would step in and correct for the noise in the signal. Essentially, rather than semantics being part of the word recognition process itself, this account would posit that semantic information was only recruited *after* (the failure of) word recognition (cf., McClelland & Elman, 1986). However, after mathematically combining EEG and MEG (for high temporal

resolution) with structural MRI scans (for high spatial resolution), it was determined that this second account was not consistent with the neuroscientific data. The behavioural data were therefore *not* the result of some decision process correcting for a token perceived as a noisy exemplar of /d/ or /t/, and semantic information *was* recruited during word recognition.

Literature support (3)

The final reason that the DCM will be favoured in this thesis is that it has been formally incorporated into accounts of word learning (Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Lindsay & Gaskell, 2010). Furthermore, it presents no inconsistencies with the lexical configuration and engagement theoretical framework (Leach & Samuel, 2007). It is also consistent with the complementary learnings systems model of memory (McClelland et al., 1995), which has likewise found favour in the literature (Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Lindsay & Gaskell, 2010).

2.1.2 A summary of the distributed cohort model

Structure of the distributed cohort model

Unlike localist models, the DCM has relatively little formal structure (Gaskell & Marslen-Wilson, 1997). Central to the model is the idea that activation of any representation is driven by direct mappings from fundamental perceptual information, such as auditory frequencies. Processing occurs without intermediate layers, for example, of acoustic signal, then mapping to phonemes, then mapping to lexemes. Highly structured, modular and hierarchical 'stages' of processing in word recognition have been seen as an outmoded oversimplification, not supported by current evidence (Gow & Olson, 2015; Spivey, 2016).

According to the DCM, lexical representations are multi-dimensional, and distributed across a network. Processing takes place such that activation of *part* of the representation then results in activation of the *whole* (as denoted by the arrow connecting lexical phonology and lexical semantics in Fig. 2.1, p. 10). Whilst the system is therefore capable of recognising sub-lexical representations (e.g., phonemes within a lexeme), this occurs by means of separate patterns of activation within the larger pattern, and not at discrete and prior stages of processing.

Semantics are also activated only according to fundamental features. The model describes the activation of semantic 'micro-features', which are summated into a data array. For example, consider a hypothetical micro-feature set 'is alive', 'has fur', 'has wings', 'eats meat'. For the representation CAT, one would expect the semantic vector {1, 1, 0, 1} as a cat has all of these features except for wings.

However, whilst they are discussed and conceptualised separately, in actuality, the model does *not* formalise a distinction between form and meaning processing. Words are instead modelled as a pattern of activated features, of any type. Words can therefore be conceived as multi-dimensional arrays, with each dimension representing a particular feature. For example, for the representation CAT, its array may contain ones for the features 'is an animal' or 'starts with /k/', but zeroes for the features 'drinks beer' and 'ends in /b/'. Recognition occurs as a pattern of activated

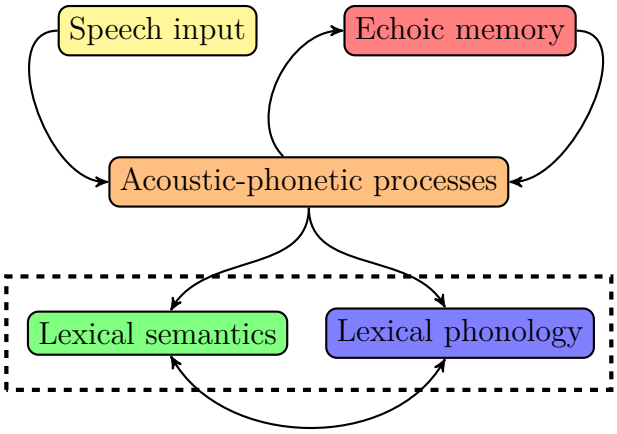


Figure 2.1: Perception of a spoken word, according to the distributed cohort model. Arrows show the flow of information processing, with the dotted rectangle showing the lexical representation. Adapted from Davis and Gaskell (2009).

nodes approaches a known pattern for a lexical representation. Using the array analogy, recognition occurs by the array moving through n dimensional feature space to a position known by the system (Gaskell & Marslen-Wilson, 1997).

Workings of the distributed cohort model

The DCM features four assumptions (Gaskell & Marslen-Wilson, 1997). These are that:

1. Lexical knowledge is distributed;
2. Different aspects of lexical knowledge are accessed simultaneously and in parallel;
3. The mappings from inputs to representations is continuous and direct, and;
4. Lexical access is maximally efficient.

A consequence of these assumptions is that as soon as speech processing begins, a 'cohort' of multiple lexical candidates is activated, with each candidate matching the incoming speech signal. For example, with the auditory input /k/ and /æ/, the representations CAT and CAP would be activated, along with any other words beginning with those sounds. Each candidate's relative activation strength would depend upon the number of times the listener had experienced that word before: frequently experienced words would be activated more strongly (Kapnoula & Samuel, 2019; Magnuson et al., 2003; Rodd et al., 2016). Thus, a word such as CANTALOUPE is unlikely to be strongly activated, despite matching the input up to that point just as well as CAT. With further information provided to the system (e.g., the perception of /t/), CAP would no longer be activated; however, CAT would be identified[1].

[1]Note that earlier disambiguation is possible if vowel-consonant co-articulation is perceived

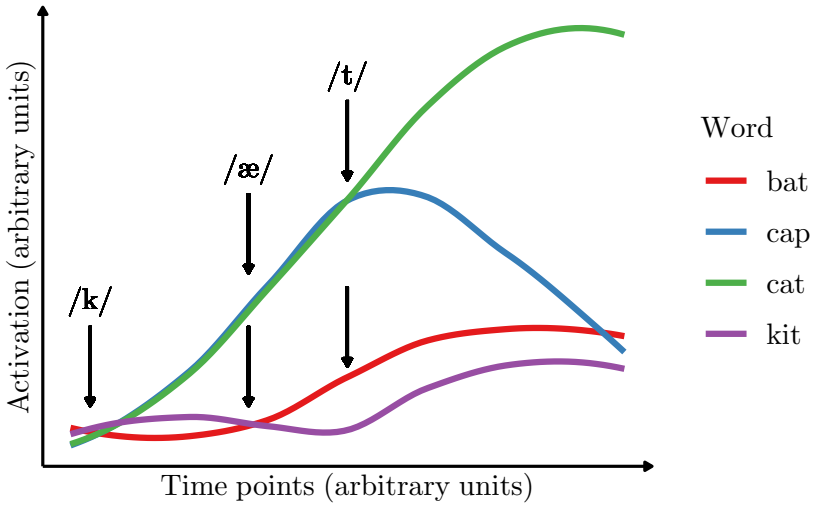


Figure 2.2: Activation of lexical representations in response to perception of /kæt/ with time, as predicted by the DCM. Note dissipation of activation for the form 'cap', and minor activation of the forms 'kit' and 'bat', in response to matching phonemes. However, lexical competition is not found for offset matched words (e.g., Dumay et al., 2004; Gaskell & Dumay, 2003; Magnuson et al., 2003). Word frequency weightings of the activation level are not depicted, for simplicity.

However, the model does *not* feature inhibitory connexions – CAP is discounted not because it is *inhibited*, but because it is no longer a good fit for the input, and its activation must therefore be weaker than the best-matching lexical candidate. The model then predicts activation of CAP to decay, as illustrated in Fig. 2.2 (p. 11). This is a consequence of the model's direct mappings (Gaskell & Marslen-Wilson, 1997).

The point at which two items can be disambiguated is called the *disambiguation point*, and the point at which a representation is in a cohort of one is its *uniqueness point*. These points may be the same, or different. For example, the representations CAT and CAP are *disambiguated* at the final phoneme. However, even at this point, they are not unique (i.e., due to items CATAPULT and CAPTAIN).

The model also explains the finding that words with a higher 'neighbourhood density', that is to say, words which have much overlap with other words, are slower to be recognised (for spoken words; Luce & Pisoni, 1998; and for written words; Carreiras, Perea & Grainger, 1997). Whilst in some models (e.g., McClelland & Elman, 1986) this slowness is explained by the time required to inhibit other units activated in the same layer, in a non-localist account, there is no single unit to inhibit. Instead, the observed slowness is explained in terms of interference. As words are represented by *patterns* of activated units, words sharing features also have similar patterning. With the activation of multiple lexical candidates, the cognitive system has more difficulty resolving the resulting interference into a detectable pattern (Gaskell & Marslen-Wilson, 1997).

In summary, according to the DCM, word recognition is a probabilistic and continuous process. From the perception of the first phoneme, input is directly mapped

to matching candidate representations, which form a cohort. The activation pattern is initially that of a 'lexical blend', composed of multiple interfering representations whose activation level is proportionate to their their relative frequency. As more input is heard, this 'blend' settles, and the activation of non-matching candidates dissipates, whilst the activation of matching candidates strengthens (see Fig. 2.2, p. 11).

In parallel to the activation of representations by phonology, semantic 'micro-feature' nodes of the distributed representations are also activated (Gaskell & Marslen-Wilson, 1997), aiding in establishing a pattern of activation matching a known representation (Gow & Olson, 2015). Recognition is complete when the system has coded for a known pattern of activation.

How then is a new word identified, without an established representation to map to? Consider the novel word 'aliet' (/eɪliət/, by analogy with 'alien', /eɪliən/). As input comes in, recognition systems would be mapping first /e/, then /eɪ/, and so on, to word representations beginning with those phonemes, such as ALE, ALIEN etc., forming a cohort of items activated according to their frequency. However, by the time the full utterance /eɪliət/ is perceived, the final /t/ acts to discount the most likely target up until that point, ALIEN. Instead, this novel word establishes a new pattern of activated units, and is recognised as such. Now encoded, this novel representation is processed further during storage.

2.2 Storing a novel word: the emergence of word-like properties

A behavioural signature of words is their dynamic interaction (Davis & Gaskell, 2009; Gaskell & Dumay, 2003; Magnuson et al., 2003), termed 'lexical engagement' (Leach & Samuel, 2007). Additionally, however, memory systems must also store static factual information, and be correctly 'configured' (Leach & Samuel, 2007). For example, words compete with one another phonologically for recognition ('lexical competition', one form of lexical engagement), but objects must also be recognised, spelling of forms must be learnt, etc., which is lexical configuration. Behavioural data concerning the emergence of such properties have led to the support for models proposing two 'minimally interacting' memory systems (Davis & Gaskell, 2009; Goldinger, 2007; Lindsay & Gaskell, 2010; McClelland et al., 1995).

One model of memory is the complementary learning systems model (CLSM; Kumaran et al., 2016; McClelland et al., 1995; McClelland, 2013; McClelland et al., 2020). Its central prediction is that episodic memory traces are *consolidated* into long term semantic memory (cf., Tulving, 1984). The CLSM has found support amongst word learning researchers, particularly those authors arguing that word learning relies on domain-general processing (e.g., Gaskell & Ellis, 2009). Furthermore, it has been argued that *lexicalisation*, the process by which words become 'word-like' (e.g., acquire their dynamic properties and become capable of engagement), is a function of the consolidation of an initial episodic memory trace of a word learning episode (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010).

It is worth mentioning that a model of word learning putting the CLSM at

its centre has been controversial. First, some authors have provided evidence apparently showing lexical engagement effects without consolidation (for reviews, see McMurray et al., 2017; Palma & Titone, 2020), suggesting that consolidation does not completely account for lexicalisation. Beyond this, there are data suggesting that representations apparently engaging in lexical engagement still have episodic properties (e.g., Kapnoula & Samuel, 2019; Qiao, Forster & Witzel, 2009). If true, these data would imply that the distinction between episodic and abstracted representation advanced by the CLSM is out-dated (e.g., Kapnoula & Samuel, 2019). Lastly, there are also models that argue that speech perception, production and lexical access are entirely episodic (Goldinger, 1998; Hintzman, 1986, 1988). However, given how influential the CLSM has been in much research (e.g., Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Goldinger, 2007; Lindsay & Gaskell, 2010; McMurray et al., 2017; Palma & Titone, 2020), and its ability to account for much of the experimental data, a summary of the model will be presented here.

2.2.1 Structure of the complementary learning systems model

The DCM represents words as distributed patterns of activated featural units. What distinguishes novel words from familiar words is that novel words are not recognised as a stored pattern by the cognitive system. Putatively, featural units are thought to correspond to neurones across the cortex (e.g., Davis & Gaskell, 2009; Hickok & Poeppel, 2007). The system of organised and distributed cortical representations is the first of the two 'complementary' systems the CLSM depends on. This cortical store is abstract, generalised, and designed for the long term storage of information and relevant inter-connexions (McClelland et al., 1995; McClelland, 2013; McClelland et al., 2020). Furthermore, it is argued that only cortical representations are capable of lexical engagement (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010).

However, the model also asserts that these representations are not 'read-only', but in some way malleable, with memory being an active and reconstructive phenomenon. The cortical store is not of immutable memories accessed passively, but rather of memories processed further at each and every retrieval. According to this perspective, memory in general, and word learning in particular, is an iterative process which may take place over months and years (Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995, though see McMurray et al., 2017; Palma & Titone, 2020).

The model argues that memories are organised into networks, and the connexions between and within representations are adjusted adaptively, although no single adjustment will be large enough to have a behavioural effect (McClelland et al., 1995). This makes the network, with its many inter-connexions, somewhat fragile, and in need of a place to store specific non-generalised, non-abstracted representations before they can be included into these delicate multi-faceted cortical representations. To resolve this 'stability/plasticity dilemma' (Carpenter & Grossberg, 1988), the model utilises a second store, containing 'episodic' (see Tulving, 1984) memories, which allows for the rapid storage of information for later inclusion into cortical networks. Moreover, this 'siloing' of information for the cortex prevents *catastrophic interference* – the wholescale destruction of the highly structured network if inform-

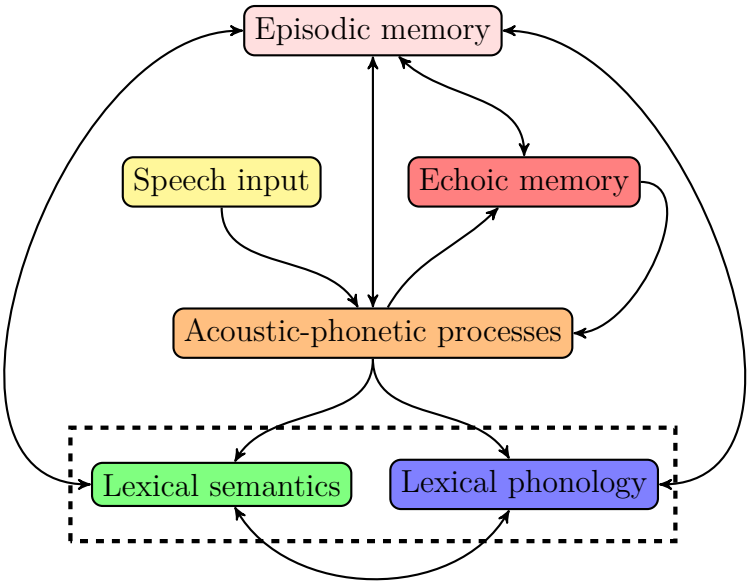


Figure 2.3: Schematic showing the inclusion of episodic memory, centred on the hippocampus, into the cortical areas utilised by the DCM (Gaskell & Marslen-Wilson, 1997; McClelland et al., 1995). As before, arrows show the flow of information during processing, and the dotted rectangle denotes lexical representations (though see Goldinger, 1998; Kapnoula & Samuel, 2019). Adapted from Davis and Gaskell (2009).

ation is incorporated too rapidly (McClelland et al., 1995; Merhav, Karni & Gilboa, 2014). This need to incorporate information gradually and iteratively explains much of the word learning data (for reviews, see Davis & Gaskell, 2009; Lindsay & Gaskell, 2010).

Episodic memory is said to be focussed on hippocampal and parahippocampal areas, following patient evidence (e.g., Gabrieli, Cohen & Corkin, 1988; Scoville & Milner, 1957). How the episodic memory system interacts with the cortical memory system during processing of a word is shown in Fig. 2.3 (p. 14).

McClelland et al. (1995) state that the pathways linking these two complementary systems – one for rapid acquisition of specific episodic detail, and one for storing in the long-term generalised, abstracted representations – are bi-directional. They argue further that, following encoding and its representation in the cortex, the novel word is next compressed and sparsely communicated to the episodic store, along with all the details from the episode during which it was encountered, including those not relevant to its productive use. Then, with time, and further encounters, the memory trace of the novel word representation is consolidated back into existing networks. The specifics of how the systems do this are outlined below.

2.2.2 Workings of the complementary learning systems model

With all the excess episodic detail, the first step of processing a stored representation is to compress it. Whilst the CLSM does not specify how the system does this, the model does explicitly reject the idea that a full copy of the memory trace is

transferred from the cortex, to the hippocampus, and then consolidated back into the cortex (McClelland et al., 1995). The authors instead argue that hippocampal representations are sparser, as certain extraneous information can be reconstructed on the fly. For example, whilst visiting Paris, one could be aware of being in France, but one does not explicitly need to encode 'I am in Paris, the capital of France' if one can rely on one's semantic memory to fill in this detail (cf., Tulving, 1984). One possibility is that the episodic (hippocampal) representation is merely a list of pointers to information stored cortically, and that the hippocampus essentially acts as a binder, or index, of distributed cortical representations (Teyler & DiScenna, 1986).

This role for the hippocampus had been well-established for almost forty years before the publication of the CLSM, on the basis of patient data (Gabrieli et al., 1988; Scoville & Milner, 1957). Following bilateral ablation of his hippocampi and parahippocampal tissues, patient HM lost the ability to form new explicit memories, suggesting that the hippocampus was critical in their formation. However, beyond simply implicating the hippocampus, the CLSM described a clear function for it. With its compressed representations, the hippocampus was modelled to act as a 'teacher' of the cortex, 'replaying' episodes to it (a process called *reinstatement*), and binding disparate elements of semantic knowledge. The cortex was then thought to organise, generalise, and abstract information across these replayed episodes (McClelland et al., 1995). In the CLSM, memories are stored as overlapping representations, prompting their co-activation on the basis of shared features, and the resulting interference during lexical processing (Davis & Gaskell, 2009; Gaskell & Marslen-Wilson, 1997; Gaskell & Dumay, 2003; Lindsay & Gaskell, 2010).

Over the longer term, the cognitive system achieves a point of full integration between novel and known information as a representation becomes less episodic in nature and is *consolidated* into the cortex. In this case, the disparate parts of the cortical representation would be independently associated, without the need for hippocampal mediation or binding (McClelland et al., 1995). Throughout consolidation, the hippocampal trace decays, as the cortical representation strengthens, resulting in a graduated 'switch' from behaviour driven by episodic representations, to behaviour driven by cortical, fully lexicalised representations (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). It has been argued that a single night of sleep is sufficient for this process to occur (Davis & Gaskell, 2009; Dumay & Gaskell, 2007, 2012). However, some experiments and authors suggest the process unfolds over longer or shorter time frames, and may be task or training dependent (Coutanche & Thompson-Schill, 2014; Hawkins & Rastle, 2016; Leach & Samuel, 2007; McMurray et al., 2017; Palma & Titone, 2020). Put another way, with or without some amount of time and possibly sleep, partial reactivation of an episode should spontaneously and automatically lead to the activation of all relevant details, now stored as a cortical representation.

This new representation will include relevant older knowledge. The system determines what is relevant by its frequency – 'dog' (/dɐg/) is mapped to the right object by the frequent co-occurrence of form and object across many experiences (McClelland et al., 1995; see also Hawkins & Rastle, 2016, for example data in a word learning study). Having learnt the word 'dog', the lexical representation

DOG should become activated either with the input /dɐg/, or by seeing the animal. Likewise, separate, already-known representations with some overlap, e.g., DONKEY (/dɐŋkiː/), should dynamically interact with DOG as it is processed (e.g., Gaskell & Marslen-Wilson, 1997). Therefore, through reinstatement, the continual interleaving of new and old experiences and knowledge, representations become consolidated. A representation is consolidated such that any part of the representation may activate the *whole* representation, and such that overlapping representations are activated below threshold by a pattern of spreading activation (Davis & Gaskell, 2009; Gaskell & Marslen-Wilson, 1997; Lindsay & Gaskell, 2010; McClelland et al., 1995, see Fig. 2.2, p. 11).

However, it should be noted that the 'switch' from hippocampally-driven behaviour to cortically-driven behaviour is not thought to be discrete, despite the observation of overnight behavioural changes (e.g., Dumay & Gaskell, 2007, 2012). McClelland et al. (1995) argue very strongly against this, whilst also rejecting the possibility of two representations and a 'dual storage' account, with each system containing its own representation and no consolidation occurring. Without a consolidation-like process to stabilise and support a memory trace, that trace would decay over time – regardless of where it is stored. However, contrary to this prediction, McClelland and colleagues point out that memory performance after some time can actually improve (in animal studies; Kim & Fanselow, 1992; Winocur, 1990; Zola-Morgan & Squire, 1990; and human word learning, e.g., Dumay & Gaskell, 2007). It is argued that this occurs because of the consolidation of a hippocampal trace, and if the hippocampus is lesioned before consolidation can occur, it cannot support consolidation of the trace into the cortex. Similar data have also been observed in patterns of amnesia caused by electro-convulsive therapy in humans, and explained in the same way (Gabrieli et al., 1988; Scoville & Milner, 1957; Squire & Cohen, 1979).

2.3 Retrieving a novel word – how is learning assessed?

Once stored, a word must be accessible, in order to be spoken or recognised. Whilst a full review of these retrieval and production processes are well beyond the scope of this thesis, it is worth reviewing how retrieval is assessed. This final section of Chapter 2 will preface a review of experimental word learning in Chapter 3 (p. 21).

In their short review of the word learning literature, Leach and Samuel (2007) distinguish between lexical configuration and engagement. They emphasise that these different ways of assessing word learning are likely to rely on different processing streams, and consequently, should be measured with different tasks.

2.3.1 Measures of lexical configuration

Lexical configuration measures – for example, referent recognition – rely on explicit and declarative knowledge (e.g., Cabeza, Kapur, Craik, Houle & Tulving, 1997). In the memory literature, measures of lexical configuration may be distinguished by whether they rely on *recognition* memory – where a participant is provisioned with some information (e.g., a form) and must use it to complete the task (e.g., matching

it to an appropriate novel referent) – or *recall*, where a participant must summon the information on his/her own (perhaps in the presence of a cue, e.g., a form's initial phoneme). In all cases, however, this explicit and declarative remembering of some aspect of a word is considered to be lexical configuration, and study designs will sometimes incorporate measures of both recognition and recall. A commonly used task is the two alternative forced choice (2-AFC; Fechner, 1860/1966), where participants must discriminate between two response options when provided with a stimulus (e.g., two referents and a form). Typically, both referents will have been learnt, but only one will have previously been associated with the form; the second referent acts as a foil. More referents may be used to make the task more difficult (e.g., 3-AFC; Coutanche & Thompson-Schill, 2014).

In the period immediately following word learning participants are often able to use their declarative knowledge explicitly to perform well on lexical configuration tasks (e.g., Dumay et al., 2004 show performance nearing ceiling on a 2-AFC task in two separate experiments). Authors who accept the tenets of the CLSM would argue that the representations supporting such performance are episodic, and that these representations are not capable of lexical engagement (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). Supporting their argument is evidence showing performance increases on configuration measures over time, suggesting that the representations are stabilising and/or being consolidated (e.g., Gaskell & Dumay, 2003; Dumay et al., 2004; Dumay & Gaskell, 2007, 2012).

2.3.2 Measures of lexical engagement

By contrast, however, lexical engagement is only ever thought to be driven by lexicalised (i.e., abstracted, non-episodic, cortical) representations. This view is held on the basis that lexical engagement is most often found for novel words following a period during which offline consolidation is thought to have occurred (e.g., during sleep; Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010). Another difference is that engagement is measured implicitly, as the measure is predicated on participants being unable to, for example, effectively manage cognitive resources during lexical competition.

The term 'consolidation' has specific theoretical implications associated with how memory traces are handled by different memory systems. However, whatever may be the case with these traces and systems, 'lexicalisation' is the process that brings about those word-like properties that make a word capable of lexical engagement. Measures of lexical engagement must therefore indirectly index the degree to which lexicalisation has occurred. As the ability to perform lexical engagement is a defining characteristic of words (e.g., Gaskell & Marslen-Wilson, 1997; Gaskell & Dumay, 2003), it may also be used as a proxy for how truly 'word-like' a learnt novel form is, and therefore how successful word learning has been. Consequently, for modelling how a form may become sufficiently word-like, lexical engagement measures yield the more theoretically interesting data. Accordingly, careful attention should be paid to the quality of the lexical engagement measure in experiments, and it is worth discussing such measures in more detail.

A distinction may be drawn between so-called 'offline' and 'online' measures of

lexical engagement. An *offline* measurement will require some processing of a word, and then the result of that process to be handed over to a decision process, which executes a response. Essentially, with an offline measure, the data collected represent serial processing of the task demands and the words themselves. By contrast, online measures may be thought to result from parallel processing of task demands and experimental stimuli, as lexical processing occurs concurrent to responding. Online measures therefore have an advantage over offline measures insofar as they allow the lexical processing to be imaged as it occurs during responding.

Offline measures of lexical engagement

One example of a commonly used offline task, sensitive to the overall level of lexical activity, is pause detection (Mattys & Clark, 2002). When performing this task, participants must press a button to detect the presence of a short (e.g., 200 milliseconds) pause inserted into a word (denoted by an underscore, e.g., 'cathed_ral'). A participant's accuracy and response time (RT) are then measured. The task itself is irrelevant, other than perhaps to measure concurrently a participant's motivation to participate. As a vehicle, however, it allows the experimenter to indirectly measure participants' lexical processing whilst, ostensibly, they perform an unrelated task.

Pause detection is applied to word learning as it measures *lexical competition* – the pattern of interference between lexical representations predicted by the DCM (Gaskell & Marslen-Wilson, 1997). The detection of a pause requires processing resources be dedicated to monitoring for it. However, when lexical competition is high, resources must be shifted away from monitoring towards the resolution of lexical competition. Performance on the pause detection task therefore suffers in environments with high lexical competition, slowing the RT. A participant's RT on the task is therefore reflective of the overall level of lexical activity in their cognitive space (Mattys & Clark, 2002; Gaskell & Dumay, 2003; Luce & Pisoni, 1998).

Referring to a lexical competition effect as the "the clearest demonstration of the lexical nature of a novel memory trace" (p. 107), Gaskell and Dumay (2003) were the first to demonstrate it between newly-learnt forms and known words. They present their data as a "stringent test of lexicalisation because it involves an effect on processing of existing lexical items", and argued "[w]hile changes in processing for novel items could have either a lexical or non-lexical locus, it is hard to argue against a lexical storage of novel sequences [if it results in] changes in the processing of existing lexical items" (p. 108). The authors taught participants novel competitors introduced at the uniqueness point of a familiar word (e.g., 'cathedruke', /kəθiːdɹuːk/, for 'cathedral', /kəθiːdɹəl/), and then measured responses to the familiar word, to gauge the impact of learning the novel word.

Their data showed that novel forms were found to move the uniqueness point of known words down the speech stream (see Fig. 2.4, p. 19), in a clear case of competition between novel and familiar words. Crucially, however, competition was only observed if a novel word had been appropriately lexicalised, and integrated with its known word competitor in the same lexical store. This took at least four days to emerge – with training across five days, and testing from the second day, no competition observed until the fourth day, despite 2-AFC performance above 90%

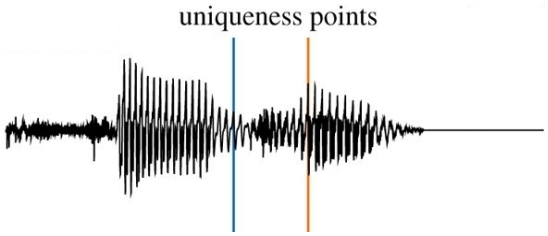


Figure 2.4: Waveform illustration of the spoken word 'cathedral', illustrating how the introduction of a novel competitor word 'cathedruke' may delay recognition by moving the uniqueness point of 'cathedral' further into the speech stream (compare the position of the blue line, before learning, and orange line, after learning). Adapted from Davis and Gaskell (2009).

on the second and third days (Gaskell & Dumay, 2003, p. 115, Table 2). Moreover, this did not occur for offset matched competitors (e.g. 'yothedral', /jɐθiːdɹəl/), eliminating some explanation based on priming, and supporting the predictions of the DCM (Gaskell & Marslen-Wilson, 1997).

Online measures of lexical engagement.

Online and offline measures of lexical engagement use the same underpinning logic, despite differences in when the task demands are processed. They are also thought to access the same (lexical) representations (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). Online measures may still look for a lexical competition effect, and index the degree to which a novel form has been lexicalised.

One notable example of an online task is the 'visual world paradigm' (VWP), an eye tracking measure (for review, see Huettig, Rommers & Meyer, 2011). Eye tracking is a very useful technique, as it very sensitive to even small changes in lexical activation driven by very newly-learnt words, and has posed challenges to existing models of word learning (Bartolotti & Marian, 2012; Kapnoula, Packard, Gupta & McMurray, 2015; Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Magnuson et al., 2003; Weighall et al., 2017). Eye tracking measures saccades and fixations. Saccades are fast, automatic, all-or-nothing ballistic movements launched around 200ms after the onset of a word (Magnuson et al., 2003; Spivey, Grosjean & Knoblich, 2005; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995); fixations are the eyes resting on an object. The assumption is that the object being fixated upon is the focus of the participants' attention and processing. With additional supporting evidence from other tasks, eye tracking has led to a re-conceptualisation of how lexical representations may emerge, and their nature in the immediate period following learning (Bartolotti & Marian, 2012; Leach & Samuel, 2007; Lindsay & Gaskell, 2013; McMurray et al., 2017; Palma & Titone, 2020).

There are four key papers in the word learning literature using VWP: Kapnoula et al. (2015); Kapnoula and McMurray (2016a); Kapnoula and Samuel (2019) and Weighall et al. (2017). As these papers all produced notable findings, they will be outlined in more depth in Chapter 7 (p. 83). However, in brief, all of them report evidence of immediate lexical competition, not predicted by previous models of word

learning (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). Furthermore, Kapnoula and Samuel (2019) present evidence not just against the time course of lexicalisation described by the original conceptualisation of the CLSM (McClelland et al., 1995), but further undermine that account by suggesting that representations capable of engagement still feature episodic details (cf., Goldinger, 1998). That is to say, the representations studied by Kapnoula and Samuel (2019) do *not* appear to be (fully) abstracted. The implications of the VWP findings will be central to this thesis.

Such findings also serve to underline the importance of *online* measures. There are various ways of constructing VWP experiments, but the data is usually analysed on the basis of the proportion of fixations to an object of interest in the visual scene. For example, during testing, participants may be presented with a screen displaying four objects. Starting at a fixation dot in the middle of the screen, participants have to, with a computer mouse, click on a referent object for a heard word. As they do this, their proportion of fixations to the objects on-screen is measured. As they survey and reject possible referents in the visual scene, participants re-orientate their attention appropriately away from distracting or competing objects, and towards their target object. A distractor object will have no overlap, semantic or phonological, with the target, whereas the competing object will compete in some way, often phonologically (for example, competition may be observed between on-screen items 'candy' and 'candle'). Essentially, this allows researchers to disentangle psycholinguistic effects from attentional ones. For example, participants fixating on a distractor object 'newspaper', in response to the input /kænd/ (for a candy/candle trial) is not likely to be evidence of a psycholinguistic effect. By contrast, the finding that participants fixate on the 'candle', when their task is to click on the 'candy', is clear evidence of lexical competition. In its various iterations, VWP has also shown novel word competition effects (between novel words, e.g., 'dibu'–'dibo', and between novel and familiar words, e.g., 'biscal'–'biscuit'; Magnuson et al., 2003; Weighall et al., 2017). More importantly, these effects have consistently been found to emerge without the need for a period of offline consolidation (McMurray et al., 2017; Palma & Titone, 2020).

With the theoretical frame of lexical configuration and engagement set out in Chapter 1, and the underpinning models of the DCM and CLSM, set out in Chapter 2, Chapter 3 will review some of the recent word learning literature.

CHAPTER

THREE

A REVIEW OF THE RECENT WORD LEARNING LITERATURE

In Chapter 2 (p. 7), a word learning model proposed by Davis and Gaskell (2009) and Lindsay and Gaskell (2010) was presented. This model of word learning unites the distributed cohort model of speech perception (DCM; Gaskell & Marslen-Wilson, 1997), and the complementary learning systems model of learning and memory (CLSM; McClelland et al., 1995). In summary, this account of word learning argues that a novel word is first encoded from fundamental mappings across cortical units, then compressed into episodic memory in the hippocampus, before being consolidated back into the cortex through reinstatement. At the point of consolidation, the novel word may then lexically engage other known words, delaying their recognition, as discussed in Section 2.3 (p. 16). This unification of the DCM and the CLSM has been influential in the literature, accounting for many experimental findings (Davis & Gaskell, 2009; Gaskell & Ellis, 2009; Lindsay & Gaskell, 2010; McMurray et al., 2017; Palma & Titone, 2020). Chapter 3 will discuss some work critical to the development of this complementary learning systems account, and summarise some questions which remain outstanding. Each question will be dealt with in turn, in its own section, but in brief, these are:

1. Is lexical engagement diagnostic of a word's lexical status?
2. What factors promote consolidation, and what is its time course?
3. What effect does the provision of semantic information have?
4. Do different learning environments produce qualitative differences in word learning, or effect its time course?

3.1 Is lexical engagement diagnostic?

This question is motivated by the fact that strong claims are made that two memory traces do not just become associated, with a novel competitor somehow pinned to a familiar word (e.g., Davis & Gaskell, 2009; Dumay & Gaskell, 2007, 2012; Lindsay

& Gaskell, 2010). Instead, it is argued that there is translation of a representation from non-lexical to lexical, represented behaviourally as increasingly generalised linguistic behaviour. In other words, rather than a mechanism based on cueing, lexical engagement is always purportedly the interaction between two 'word-like' traces (cf., Pufahl & Samuel, 2014). The current section interrogates this claim a little further.

3.1.1 The theory for minimally interacting complementary systems

The first challenge to CLSM is the relatively slow, or unreliable, emergence of lexical engagement (compare Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Weighall et al., 2017). Clearly, once learnt, the representation is present in the mind immediately. Why then is it that this representation may drive performance on lexical configuration, but not lexical engagement, tasks? From a CLSM perspective, with two systems supporting behaviour, one must explain precisely how and when each system is acting, and why they do not interact.

Davis and Gaskell (2009) propose a neat solution to this problem. They predict that, following consolidation (purported to drive lexicalisation), a target word 'captive' (/kæptɪv/), a familiar competitor 'captain' (/kæptən/), and a novel competitor 'kaptik' (/kæptɪk/) will all be part of the same lexical blend when the systems perceives the input /kæpt—/. Recognition of 'captive' will then be appropriately slowed (e.g., Luce & Pisoni, 1998). In this case, the novel representation is said to be stored cortically. This is a simple restatement of the DCM, including a newly-learnt word.

The situation *prior* to consolidation is, however, more complicated. In this case, Davis and Gaskell argue 'kaptik' should not be free to engage 'captive', due to the "isolation of the hippocampal route, [meaning] that the relative probability of [kaptik] cannot be properly incorporated into the weighted [lexical] blend" (p. 3779). This initial failure of 'proper incorporation' is driven by a weighting of the hippocampal and cortical processing routes. However, this creates a further problem, as a cognitive system tuned too heavily towards the hippocampal route would cause catastrophic interference of memory traces, where the clumsy overlapping of representations makes all representations non-readable (McClelland et al., 1995; Merhav et al., 2014). By contrast, a system tuned against the hippocampal route would struggle to recall any novel information before sleep.

The solution presented is to factor in a 'dominance' of the cortical processing route up until the failure of recognition processes, at which point, the cognitive system could read episodic (i.e., hippocampal) memory to find a potential match. Essentially, Davis and Gaskell clarify that the third and fourth assumptions of the DCM (that mappings are continuous and maximally efficient, p. 10) should only apply to those representations thought to have been 'consolidated'.

3.1.2 The possibility of experimental artefacts

Accepting that there are distinct stores, and that these separately contain different representations, a second challenge frames reported lexical competition effects as experimental artefacts, whereby non-lexical traces disrupt lexical processing. Lexical

competition effects with novel words can thus not be said to be due to the lexicalisation of a novel word, and lexicalisation must therefore occur under different conditions. Supporting this idea is the finding that novel words may disrupt each other without lexicalisation (i.e., competition between newly learnt forms 'dibu' and 'dibo'; Magnuson et al., 2003). This implies that competition *generally* may not be a property of only lexical items, and that the 'type' of competition should be considered: it is only lexical where it occurs between *lexical* items (Gaskell & Dumay, 2003; Pufahl & Samuel, 2014). Moreover, it has also been shown by Magnuson et al. (2003) that novel forms can immediately possess other word-like properties (e.g., sensitivity to frequency of occurrence; neighbourhood density effects – see Luce & Pisoni, 1998). It therefore seemed possible that novel forms would be difficult to distinguish from familiar words experimentally, and conceivable that effects attributed to lexicalised forms were in fact driven by non-lexical representations (not accounting for the theoretical dominance of the cortical route).

Qiao et al. (2009) presented an example of this second challenge. The authors were responding to Bowers, Davis and Hanley (2005), who presented evidence of (orthographic) novel lexical competition effects. Recognising that lexical findings may be contaminated by lexical properties such as frequency of occurrence, Bowers and colleagues drew up a list of 'hermit' words (words for which there are no orthographic neighbours, e.g., 'sleeve', 'banana'). Orthographic neighbours were words which could be created from another word with the addition, deletion or replacement of a single letter (e.g., 'sleeve' → 'sleere', by replacement). From these hermits, Bowers and colleagues created a series of replacement neighbours. This provided the authors with a sufficiently well-controlled stimuli list for them to investigate the impact of novel word learning 'cleanly', without problematic confounds from other lexical properties. They used a semantic categorisation task as their engagement measure. In this task, participants made speeded categorisation judgements: participants had to identify the familiar words as either artefacts, or natural objects. This was felt to better eliminate the possibility of episodic (phonological/orthographic) cueing than a task requiring a judgement about the word itself (e.g., lexical decision, which requires participants to make a word/non-word judgement for each trial stimulus). Semantic cueing was not a concern as the novel words were trained without semantic referents. Comparisons were made between two lists of items, for one of which, novel competitors had been trained. It was anticipated that a lexical competition effect would emerge for the list with competitors, and semantic categorisation response times (RTs) would be slower for those items (cf., Dumay et al., 2004; Gaskell & Dumay, 2003).

Testing over two consecutive days, participants performed the semantic categorisation task three times: immediately after training on the first day, before a block of training (of the same novel words) on the second day, and after training on the second day. Bowers and colleagues found no difference between the RTs to current and former hermit words immediately after training on the first day, but by the second day, this effect was evident in both tasks – perhaps suggesting evidence of consolidation as the novel forms became integrated with the familiar words overnight. Interestingly, given that on the second day participants performed the engagement task before, and after, further training, it was possible to see the effect

of further training on a fragile novel trace (cf., Goldinger, 1998; Kapnoula & McMurray, 2016b). However, whilst the RT difference strengthened between all tasks, only the difference between the first and the third tasks was statistically significant.

A non-lexical basis for 'lexical' effects

As mentioned, Qiao et al. portray all the above effects as artefacts. They argue that the novel word trace may still be non-lexical, and an RT cost of processing 'banana', explained as lexical competition from 'banara' by Bowers et al., is instead driven by a need to check what one has just seen, post-lexical access. Aware of having recently learnt a word similar to 'banana', participants felt the need to verify that they had indeed seen 'banana' when it appeared on screen. This selectively slowed them down, and produced a pattern of data which, Qiao et al. argued, Bowers et al. mistook for a competition effect.

Whilst this is in some sense still lexical engagement – 'banara' has in some way become linked to 'banana' – it is not the automatic effect of more difficult lexical processing one would usually expect to see following learning, and Qiao et al.'s critique is therefore valid. They continued their line of reasoning: if lexicalisation had truly occurred, then forms would show robust lexical effects across many measures. They therefore proposed testing using a masked priming paradigm, as a more direct measure of the learnt words lexical status. Qiao et al. cited the 'prime lexicality effect', whereby a lexicalised form produces no priming or inhibition of a target, whereas a non-lexical form produces facilitation. In a masked priming task, participants have to make some judgement to a target (e.g., 'banana'), when it is preceded by either a prime, or a non-prime, and a mask (######). Qiao et al. showed a mask for 500ms, followed by a (lowercase) instance of the prime (e.g., 'banara') or non-prime (e.g., 'agenty'), followed by the (uppercase) target on screen for 500ms. The experiment had a 2 × 2 design for the novel words preceding familiar word targets: prime status (prime, non-prime) and trained status (trained, untrained). Unsure of whether Bowers and colleagues semantic categorisation task was applicable to masked priming, Qiao et al. replaced that task with a lexical decision task.

Qiao et al.'s data are problematic for Bowers et al.'s account. They found no evidence of a reduction in priming of responses to 'banana' by 'banara' across two days of testing, suggesting that the novel form remained non-lexical, according to the prime lexicality effect. Moreover, in another condition of the experiment, they were able to show that familiar words (e.g., 'passive') showed no effect of priming a related known word (e.g., 'massive'), relative to a familiar non-prime (e.g., 'logical') – suggesting the problem was not with demonstrating the prime lexicality effect generally. The data were unable to be dismissed as a function of poor learning – whilst no measure of lexical configuration was taken, a main effect of training was found, suggesting that learnt traces were processed differently to unlearnt ones and therefore that participants did maintain some representation of the trained forms. The data from Bowers and colleagues has more recently been replicated and extended (Coutanche & Thompson-Schill, 2014; Walker et al., 2019; Wang et al., 2017), but Qiao et al.'s data do need to be integrated into an account of word learning.

3.1.3 A stronger case for a lexical locus

An obvious way of resolving Qiao et al.'s criticisms would be to use less similar stimuli. A competition effect demonstrated under conditions not so similar as to trigger the proposed post-access check would likely be true evidence of lexicalisation (Davis & Gaskell, 2009; Dumay & Gaskell, 2012; Gaskell & Dumay, 2003). This work has been done as well. In two experiments comparing onset competitors (e.g., 'frenzylk', for 'frenzy') to embedding competitors ('lirmucktoze', for 'muck'), Dumay and Gaskell (2012) were able to directly examine the effect of competitor similarity. Novel onset competitors were largely composed of familiar words and thus very similar to targets (as in Gaskell & Dumay, 2003), whereas embedding competitors only had a minority of their phonemes in common with targets, and thus were quite different. Testing occurred on the same day, the next day, and a week after training. The two experiments also used two separate lexical engagement tasks and different participants. The first experiment used pause detection similarly to Gaskell and Dumay (2003). However, the second experiment used 'word spotting' – where a participant had to press a button when they detected any embedded word. Like pause detection, word spotting records an RT indicative of the difficulty of the task.

As proponents of a consolidation account, Dumay and Gaskell expected that both word spotting and pause detection would show an RT cost indicative of learning in both sets of competitors. Likewise, it was expected that this cost should only be present on or after the second day of testing. Moreover, as lexicalisation took place, it was thought that participants would cease processing embedding competitors as nonsense syllables wrapped around a 'real' word (i.e., less like 'lir**muck**toze'). Instead, embedding competitors would be processed lexically (perhaps, much how the lexical item 'badminton' does not consciously evoke 'mint'). Thus, when compared with onset competitors, where a familiar word was immediately presented to participants in the first few phonemes (i.e., **frenzy**lk), participants should show slower word spotting RTs for embedding competitors. This change was thought to occur as the saliency of the embedded word decreased with lexicalisation, and the new form 'lirmucktoze' was processed more as a lexical unit in its own right. The authors contrasted this with Qiao et al.'s non-lexical account, which Dumay and Gaskell argued would predict facilitation. This is because, during training, participants would code the position of the known word within the novel carrier, particularly where the novel form was an onset competitor[1], and that this information would be exactly that needed to perform well in the word spotting task. As Qiao et al. argued that competition effects emerge from a non-lexical representation interfering with a familiar target, across days with further consolidation, the word spotting task should have become steadily easier.

However, at test, Dumay and Gaskell showed evidence for lexicalisation over time, with competition effects on the second and seventh days of testing, but not on the first – confirming previous work (Bowers et al., 2005; Davis & Gaskell, 2009; Dumay et al., 2004; Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Lindsay & Gaskell, 2010; McClelland et al., 1995). Moreover, findings were also consistent with a consolidation account, insofar as both pause detection and word spotting showed

[1] As here the known word was more salient, as it occurred at the beginning of the word.

competition effects, not facilitation, which were larger for embedding competitors. This paper is key in supporting the contention that lexical engagement is diagnostic of lexical status.

3.1.4 The case for consolidation effects in word learning

Collectively, these theoretical accounts and empirical data present a strong case for consolidation, and lexicality emerging over time. Late-emerging competition effects are also quite common in the literature (e.g., Bowers et al., 2005; Dumay et al., 2004; Dumay & Gaskell, 2007, 2012; Gaskell & Dumay, 2003; Hawkins & Rastle, 2016; Henderson, Weighall & Gaskell, 2013; Tamminen & Gaskell, 2008; Walker et al., 2019; Wang et al., 2017). Given this, even accounts that do not accept that lexicalisation "will not be hurried" (Dumay et al., 2004, p. 344) *do* accept some sort of role for strengthening or deepening representations with consolidation (Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Lindsay & Gaskell, 2013; McMurray et al., 2017; Weighall et al., 2017). The case is further supported by competition effects being demonstrated in the long term (as McClelland et al., 1995, originally suggested, over a period of many months; Tamminen & Gaskell, 2008). Further to this, McKay, Davis, Savage and Castles (2008) demonstrated that even up to a year after learning, participants (who failed to remember the definitions of many words they had learnt) were still faster to read aloud trained novel words than reading-difficulty matched control items. It seems hard to conceive how an episodic trace would be inaccessible for explicit recall, and yet facilitate reading performance – instead, suggesting some abstraction and lexicalisation of a cortically-stored novel word. Neither is it likely that frequent reactivation of the novel word across training and testing is a likely explanation for the results – experimenters have applied controls for this, and found no effect (e.g., Dumay & Gaskell, 2007). Lastly, although there may be some qualitative differences between children and adults, evidence of consolidation during word learning has been found in children (Brown, Weighall, Henderson & Gaskell, 2012; Henderson et al., 2013; Henderson & James, 2018; Weighall et al., 2017). Therefore, in broad terms, the consolidation account can largely be accepted.

There is, however, some noisiness in the data. An obvious and glaring inconsistency is in the time frame reported for consolidation-like effects. Setting aside papers finding evidence for immediate lexical competition for later chapters, it has been reported that consolidation may happen in a single 24-hour period (Davis & Gaskell, 2009; Dumay et al., 2004; Dumay & Gaskell, 2007, 2012; Wang et al., 2017); however, some of the same authors previously reported no effect until the fourth day of training and testing (Gaskell & Dumay, 2003). Other authors have likewise found longer timescales for engagement effects to emerge (Leach & Samuel, 2007; Hawkins & Rastle, 2016; Walker et al., 2019). Children have likewise demonstrated inconsistency (although measures of configuration did show evidence of consolidation, and the authors admit to poor task design, Brown et al., 2012). Whilst the authors do not explain their inconsistencies directly, some combination of design and task differences is likely to be at work (see Kapnoula & McMurray, 2016a; Leach & Samuel, 2007; Walker et al., 2019; Weighall et al., 2017). It is therefore worth considering

what factors may promote lexical engagement.

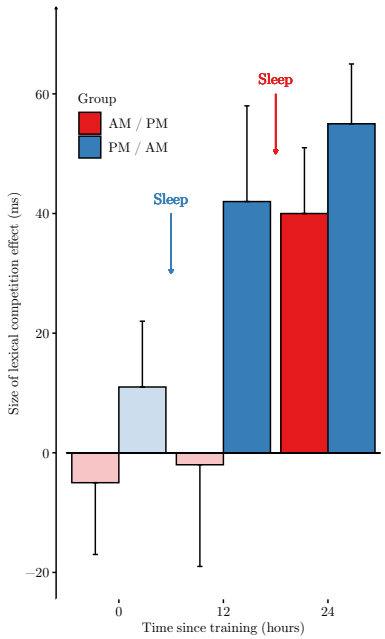
3.2 What factors promote consolidation, and what is its time course?

3.2.1 The effects of sleep and time

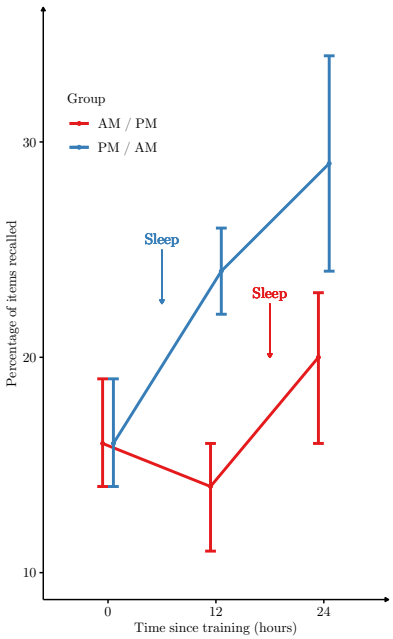
Sleep has been argued to be the most important factor promoting consolidation, acting as a mediator between the complementary systems (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). Although accepting that it is not yet possible to draw a causal link between sleep and consolidation (as a state associated with sleep, rather than sleep itself, may promote consolidation), Davis and Gaskell (2009) argued strongly for some involvement of sleep. However, whilst the state may be implicated, they did not argue that a single night of sleep would imply 'dichotomous knowledge transfer'; rather, merely that it would be sufficient for behavioural change (a conceptualisation of consolidation as an ongoing and continuous process that may also be useful in explaining some discrepancies in the emergence of competition effects). The key study for this conclusion was Dumay and Gaskell (2007, though see also Bowers et al., 2005; Brown et al., 2012; Dumay et al., 2004; Gaskell & Dumay, 2003; Tamminen & Gaskell, 2008; Tamminen, Payne, Stickgold, Wamsley & Gaskell, 2010; Wang et al., 2017; Walker et al., 2019).

An empirical challenge in word learning research has been how to disentangle the effects of time and sleep. To address this, Dumay and Gaskell compared participants trained and tested across three sessions. The three sessions were: immediately following testing, after 12 hours, and after 24 hours. At some point within this period, participants slept, and groups matched on their self reported amounts of sleep. However, by systematically varying the time at which testing began, the experimenters could look to see the effect of sleep on their lexical configuration (2-AFC, free recall) and engagement (pause detection) measures. Participants were split into two groups, and training began either in the morning, with testing then in the evening of the same day, and again the next morning (the 'AM/PM' group), or else in the evening, and then the morning of the next day, and again, later on the evening of the second day (the 'PM/AM' group).

A graphical summary of all the data can be seen in Fig. 3.1 (p. 28). Results showed that in either group, no lexical competition effect was immediately present. However, it did emerge after sleep in both groups (i.e., after 12 hours for the PM/AM group, and after 24 hours for the AM/PM group). For the PM/AM group, this effect was maintained but did not strengthen reliably throughout the course of the second day (Fig. 3.1a). Free recall performance also showed a sleep-related performance effect: whilst the AM/PM group declined between the first and second testing session, performance then recovered and improved with sleep. By contrast, the PM/AM group showed reliable increases in performance across all three sessions (Fig. 3.1b). Lastly, the 2-AFC task – where participants had to decide if a form they heard was one they learnt, or not – showed immediately near-ceiling performance, which remained stable across all three sessions. To test for reactivation of the traces driving the other effects in the study, half of the participants were only given the



(a) Pause detection data, with significant lexical competition effects shown by the saturated bars.



(b) Free recall performance.

Figure 3.1: Data from Dumay and Gaskell (2007).

2-AFC test in the final session. However, this revealed no difference in performance, and in any case, would not be sufficient to explain the differential emergence of the competition effects. Dumay and Gaskell therefore stated there were three possible explanations of their findings. These have yet to be clearly disentangled in the word learning literature, and to do so is beyond the scope of this thesis. However, they are:

1. Some state occurring with sleep promotes consolidation (possibly also related to the circadian rhythm);

2. Poverty of input during sleep allows sufficient 'downtime' for consolidation;

3. Sleep has a causal role in promoting consolidation.

Further data implicating sleep has also been found. Polysomnography, a neuroscience technique which allows for the recording of brain data whilst a participant sleeps, was used by Tamminen et al. (2010) to investigate further Dumay and Gaskell's effects. In a similar design, but with the third testing session a week later, the authors showed that particular elements of the sleep cycle mapped to behavioural performance, supporting the idea of a causative relationship between sleep and consolidation.

3.2.2 The effect of further exposure

Whilst the number of exposures during training will be considered separately (Section 3.4, p. 36), it is worth noting that the number of exposures overall is significant in bringing about lexical engagement.

Research has shown that further exposure alone does not promote lexical engagement. Finding no competition effect until the fourth day of a five-day study (Experiment 2), Gaskell and Dumay (2003; Experiment 3) again looked for a competition effect at the same three time points as in their later work and Tamminen et al. (2010). Participants in Experiment 2 had been given training every day; participants in Experiment 3 were therefore given massed training on the first day of testing equivalent to those three days in Experiment 2 without an effect. Critically, however, in Experiment 3 no training was given in the interim between testing points. Results showed that no effect again emerged until after sleep. Collectively, these data suggest that lexicalisation cannot be sped up by exposure alone and that sleep is critical (though see Lindsay & Gaskell, 2013).

However, whilst it may not be the case that exposure alone can speed up lexical competition, it is worth noting that the overall number of exposures does seem to be an important variable. This is particularly of concern with studies which use the 'fast mapping' paradigm, used later in this thesis (for review, see Cooper et al., 2019c). Much of this work has been conducted in the orthographic modality, so may not produce results which are directly comparable in terms of the number of exposures required, although the broad pattern of data may still be applicable. Recent evidence has shown that whilst conditions of high exposure (10 or 20 repetitions per word) led to detectable lexical competition post-sleep, five exposures was insufficient, despite there being evidence of consolidation in the lexical configuration measures (Walker et al., 2019). It may be that to detect lexical competition after a single night of sleep, the to-be-consolidated episodic trace must be already of sufficient strength (cf., Cooper et al., 2019c).

In summary, therefore, it seems that for a representation to engage other lexical items, it must be of sufficient strength. Two factors that may stabilise and strengthen representations, to allow the detection of lexical engagement, are the number of initial exposures and overnight sleep. However, it should also be recognised that this conclusion is reached from looking at studies which use an impoverished training regime, not representative of normal word learning (Gaskell & Dumay, 2003; Dumay et al., 2004). In particular, participants do not learn semantics (e.g., Bowers et al., 2005; Brown et al., 2012; Dumay et al., 2004; Dumay & Gaskell, 2007, 2012; Qiao et al., 2009; Tamminen & Gaskell, 2008; Tamminen et al., 2010; Walker et al., 2019; Wang et al., 2017). One possibility is that under such circumstances, without the provision of semantic information, the cognitive system struggles to integrate a novel form – perhaps because by definition, *words* are meaningful, and must necessarily include semantic information (Gaskell & Marslen-Wilson, 1997; Gow & Olson, 2015; Spivey, 2016). A forms link to lexical items may be tenuous as the system struggles to recognise it as 'word-like'. It may be that only under these circumstances are sleep and the number of exposures relevant factors. A related question is how the *type* of training influences word learning (cf., Coutanche & Thompson-Schill, 2014;

Kapnoula & Samuel, 2019; Leach & Samuel, 2007; Sharon et al., 2011). These questions have been examined by the literature, and the findings will be reported in the final two sections of this chapter.

3.3 What effect does the provision of semantic information have?

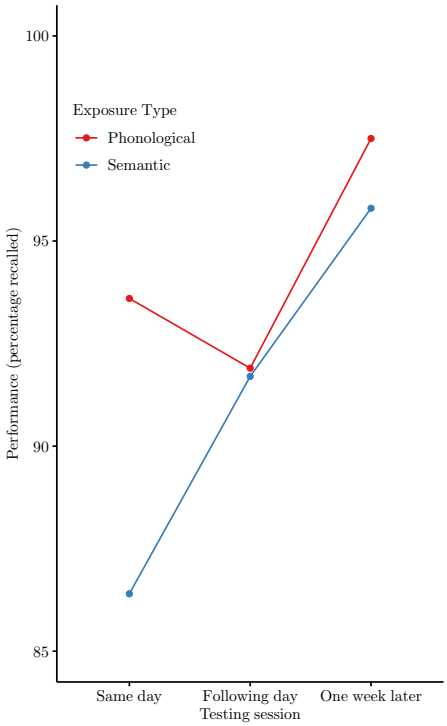
3.3.1 The case against semantics supporting word learning

Dumay et al. (2004, Experiment 1) is an example of an early study investigating semantics and competition effects. The authors compared words learnt by 'phonological exposure' (phoneme monitoring) against words learnt in a 'semantic verification' task. Phoneme monitoring required participants to listen for the presence of a particular phoneme, and make speeded judgements as to the presence or absence of that target (e.g., /k/'s presence in 'cathedruke'). Participants completed twelve blocks and therefore were exposed to each word phonologically twelve times. The semantic verification task took place in two blocks, with six sub-blocks, balancing exposure of the novel words learnt by each method. In each of the blocks, participants either heard the word in explicit and non-explicit contexts (e.g., "A 'cathedruke' is a variety of vegetable"; "The cook served the boiled 'cathedruke' with a steak and baked potatoes"). In each of the six sub-blocks, participants had to make a speeded yes/no decision about some property of the word (e.g., about its edibility).

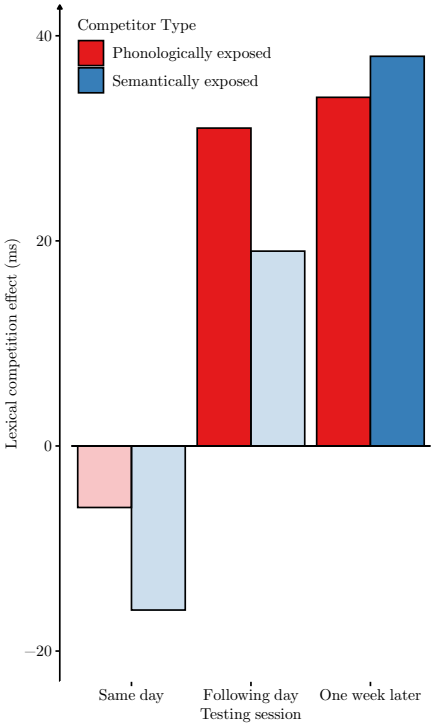
As with other work (e.g., Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008), testing took place on the same day as training, the day following training, and a week later, and configuration and engagement data were collected. Configuration was tested by 2-AFC ('cathedruke'/'cathe*druce*' discrimination). Engagement was measured with three tasks. Firstly, lexical decision was used against the base words, from which the novel competitors had been derived (e.g., 'cathedral', for 'cathedruke'). This looked for lexical competition as has already been described (see Section 2.3.2, p. 17). Secondly, a more 'semantic' task was also used. Occasionally, a lexical decision trial featured the novel word, and the next trial featured the category name (e.g., 'vegetable', preceded by 'cathedruke'). This task measured the ability of the novel word to prime its category name, and facilitation was expected. Finally, the ability of the novel words to prime other words was tested in a 'free association' task. Responses to this final task were coded as one of 'novel word meaning', 'base word', 'base word meaning', and 'other'. The findings were as follows, and are summarised in Fig. 3.2 (p. 31).

As shown in Fig. 3.2a, 2-AFC performance was good, with participants having strong declarative knowledge of the correct form. Performance was superior for phonologically-trained words in the first session, but equivalent thereafter. Both sets of words saw statistically significant strengthening of recognition ability across testing, suggesting sleep-related consolidation, as has been seen elsewhere.

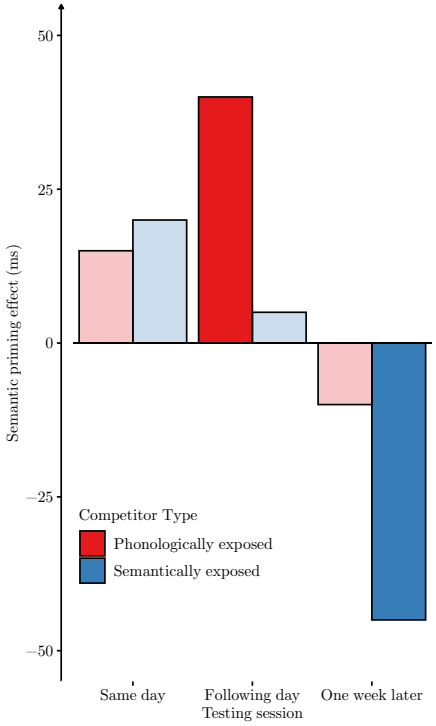
Fig. 3.2b illustrates the data from the lexical competition trials. No lexical competition effect was present in the first session for either type of novel word. However, a significant lexical competition effect was evident from the second testing session, for phonologically-trained words alone, which persisted through to session



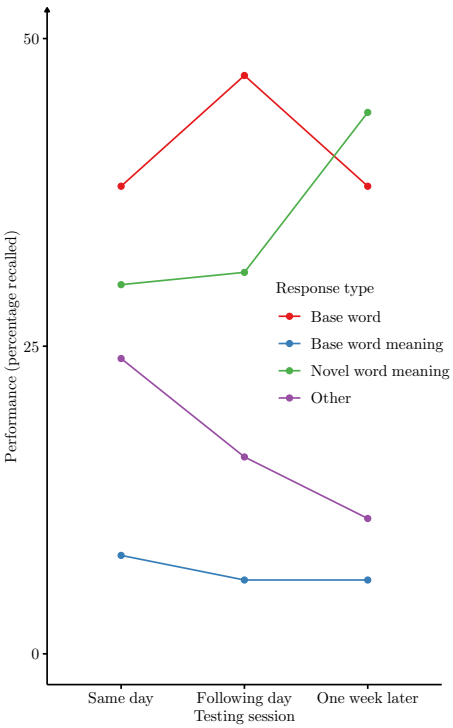
(a) 2-AFC performance



(b) Lexical competition data. Saturation indicates significance; competition in positive x direction



(c) Semantic priming data. Saturation indicates significance; facilitation in negative x direction



(d) Free association data

Figure 3.2: Data from Dumay et al. (2004, Experiment 1)

three. By contrast, semantically-trained words lagged behind – no competition was observed until a week after learning. Unfortunately, no statistical comparison of the size of the effect at this point was reported, however, it is notable that semantically-trained words produced a numerically larger effect.

It could be argued that this was an unsurprising finding. The lexical competition aspect of lexical engagement primarily measures engagement between *forms*, and therefore, perhaps words learnt in a way which drew attention to their phonology (i.e., phoneme monitoring) would produce a statistically reliable effect faster than semantic training. It may have been that use of a semantic task confers a similar advantage for semantically trained words. Fig. 3.2c shows some evidence for this. Although it was not so fast to emerge, only semantically-trained words showed evidence of priming. Moreover, responses consistently trended towards facilitation for semantically trained words, across testing sessions. Phonologically-trained words showed an effect in the *wrong* direction on the day following training; however, this disappeared after a week, and in any case, must have been noise as participants were not aware of the category for a phonologically-trained words. Phonologically-trained words in this condition simply functioned as a baseline to compare semantically-trained words to. However, it seems that if one considers lexical engagement narrowly, and only through the 'keyhole' of lexical competition, semantics do not support faster lexicalisation, and are slower to emerge.

The free association task, shown in Fig. 3.2d, allows one to draw together these two aspects of lexical representations. Whilst on the same and following day of testing, participants most often elicited the base word (e.g., 'cathedral', for 'cathedruke') – probably suggesting declarative knowledge of the novel word's form, due to the similarity (even if this does not drive the competition effects; Dumay & Gaskell, 2012) – after a week, participants most frequently referred to the novel word's meaning. Whilst in the absence of familiar word data for the free association one cannot draw firm conclusions about the lexicality of the representations driving these findings, this is nevertheless interesting, as it matches the data shown in Figs. 3.2b and 3.2c. Caution should be used, however, as it is unclear precisely what the task indexes. For example, it could be the case that the novel word automatically evoked either its competitor or category – a clear case of engagement – or that participants were reliant on knowledge stored at encoding. That 'cathedruke' sounds like 'cathedral', or that it is a vegetable, is essentially static knowledge, and therefore, configuration. Furthermore, it may even have be that the task did not rely on consistent processing across time points – for example, with a switch from configuration-like to engagement-like processing.

In the round, the data from Dumay et al. (2004) suggest that the provision of semantic information does little to support the emergence of a lexical competition effect. In turn, this suggests that cognitive systems are not discriminatory in the processing of forms, which appear to be handled as word-like *enough* to be put through lexicalisation (or a process mimicking it well enough to produce indistinguishable lexical decision effects). For semantically-trained items, it is still not the case that they show an advantage on semantic tasks, despite equivalent knowledge seen in the 2-AFC task. It seems likely, therefore, that semantic effects are slower to emerge than lexical ones.

Other studies have found similar patterns of data. For example, McKay et al. (2008) found that the provision of semantic information during training only supported the ability of children to read novel words aloud where the pronunciation was not consistent with its spelling. This appears to be some limited evidence for semantics supporting lexical processing; however, this finding may not have been semantically supported. In a third experiment conducted up to a year after training, many participants had forgotten the meanings, but the effect still persisted. Poor child readers have however shown evidence that an accompanying picture supports reading behaviour (McNeil & Johnston, 2004). Note that this underscores that 'semantic studies' often have quite different designs, which, in turn, may bring about quite different effects, but which are not attributable to having learnt semantic information, *per se*.

However, more recent data have shown that semantic learning need not be as slow as Dumay et al. (2004) first suggested. Henderson et al. (2013) again compared semantically-trained and phonologically-trained novel words learnt by children. To assuage fears that constructed novel words (e.g., 'cathedruke') may not be handled as words (e.g., Leach & Samuel, 2007; Qiao et al., 2009; Tham, Lindsay & Gaskell, 2015), the authors used real words (e.g., 'hippocampus'). The training regime was also a little different to previous work. In both conditions, participants learnt a word, and saw something (either a novel referent or the printed novel word). Also in both conditions, participants repeated and segmented the word, repeating first the whole word, then its initial syllable, and then its final syllable (see also Brown et al., 2012; Weighall et al., 2017). In the last training task, participants either made a semantic decision (was the word an animal, a plant or neither) or a phonological one (was the word composed of 2, 3, or 4 syllables). At test, competition effects were measured by pause detection, and testing took place at the usual time points (immediately, following day, one week later; cf., Dumay et al., 2004; Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008). By and large, results showed no difference between types of training, only across sessions. The pattern across sessions was however different. Whereas no competition effect emerged for either type of trained word immediately, it was also not present after a week. This may have been due to young children being unable to retain words over longer intervals without repetition, or due to noisy responding (see also Brown et al., 2012). A competition effect was present for the semantically-trained words after 24 hours, but this was marginally non-significant ($p = 0.08$) for the phonologically-trained words. Lexical configuration was also measured, with results much the same as previous studies (evidence of consolidation, strong declarative knowledge) and again, no effect of training type.

Nevertheless, this study provides evidence that the presence of semantic information need not *slow* the emergence of lexical competition effects. Other studies have made much the same arguments with quite different tasks (Hawkins & Rastle, 2016; Tham et al., 2015). Instead, these data provide evidence that processing is flexible enough to be insensitive to the absence or provision of semantics. This set of papers suggests there is neither a benefit, nor a cost, which results from semantic training.

3.3.2 The case for semantics supporting word learning

Unlike researchers supporting their arguments with phonological data, Tham et al. (2015) measured lexical engagement using semantic measures only. Their study was with English speakers learning translations, either with Chinese logographs (Experiment 1), or Malay (Experiment 2). Experiment 2 largely replicated the effects of Experiment 1, but was run to allay fears that the logographs would not be processed as words by English speakers used to writing/reading in the Latin script, which is also used by Malay.

Tham and colleagues selected two measures of semantic integration – the size congruity and the semantic distance effects. When participants have to select between two words printed on screen, the size congruity effect describes the phenomenon that RTs may be increased by an irrelevant dimension. For example, participants will be faster to identify that a cow is larger than a bee (by selecting the word 'cow' on screen) when the word 'cow' is written in physically larger font also (congruent trial). This contrasts with an incongruent trial, where the font for the word 'bee' is larger (i.e., BEE — COW). The semantic distance effect refers to the fact that this judgement is made harder or easier still by the similarity of the relevant dimension (here, the size of the animal, not the font). For example, selecting the larger animal between 'bee — dog' will be harder than selecting between 'bee — cow'. With these examples, the combination of these two effects would predict slowest RTs on a trial 'BEE — DOG', and fastest RTs on a trial 'BEE — COW'. Tham and colleagues claimed that of these two effects, size congruity is the more sensitive of the two, and better marker of lexical engagement. Testing was at three points, the time of which varied by sleep/wake group, as in Dumay and Gaskell (2007). In two experiments, Tham et al. found that the size congruity effect emerged after sleep, in line with consolidation accounts. More interesting is that of an equally reliable effect of semantic distance *before* sleep (Experiment 1: post-sleep size congruity effect Cohen's $d = 0.37$, pre-sleep semantic distance effect Cohen's $d = 0.38$; Experiment 2: post-sleep size congruity effect Cohen's $d = 0.27$, pre-sleep semantic distance effect Cohen's $d = 0.32$). Although Tham et al. argue that these data are *not* indicative of full automaticity and engagement, the finding is interesting as it hints at later findings (e.g., Coutanche & Thompson-Schill, 2014; for reviews, McMurray et al., 2017; Palma & Titone, 2020). Although Tham et al.'s findings do not sit completely neatly with the work described above (due to learning being in a second language, and the orthographic modality), it is an indication that newly-learnt words with semantic meaning need not be slow to show engagement (and indeed, may be faster than expected).

Hawkins, Astle and Rastle (2014) also demonstrated findings of note. Instead of comparing groups of novel words trained with or without semantics, they trained all of their novel words with semantics, but only half of this set were reliably associated with a particular referent, whilst the rest had an equal chance of appearing alongside any referent. This standardised training entirely but still allowed the researchers to look the effects of learning a referent. At test, rather than looking for lexical engagement, Hawkins et al. tested for knowledge of the learnt words form. Instead of asking participants to explicitly make a judgement, however, the research-

ers measured knowledge of the form by attempting to elicit a mismatch negativity (MMN) event-related potential (ERP). An ERP is a neuroscientific technique measuring brain activity in response to a specific cognitive event, and the MMN occurs when participants, habituated to a stimulus, hear a 'deviant' stimulus. The observed activity is indicative of the activation of a memory trace for the deviant stimulus. In Hawkins and colleagues' experiment, at test, participants received 900 presentations in total, of which 300 presentations were deviant ('boap' or 'boak'), and half of these deviant stimuli had been reliably associated with a referent (e.g., BOAP). The remaining 600 were a familiar word (e.g., 'boat'). The data showed larger MMNs for forms reliably associated with a referent. Moreover, as the experiment tested over two days, on the first day, the size of the MMN was found to be associated with the accuracy during training – suggesting better learning meant a stronger MMN. However, this was not the case on the second day – perhaps as the form became consolidated and stabilised. Although these data concern only knowledge of a form, and thus do not testify as to how semantics may or may not support lexical engagement directly, the fact that some behaviours at least are supported by semantics may leave open the possibility of improving lexical engagement with semantic training.

This theme has been recently built upon. In the context of word learning in either one's first language, or a second, Havas et al. (2018) trained participants with either known or novel referents. Participants performed two tests of recognition memory: a 4-AFC and a new/old discrimination task, where participants had to press one of two buttons to decide if a heard word had been previously trained. The researchers found that, in either language context, where a participant learnt another word for a known object (i.e., equivalent to learning either a synonym or its translation, instead of learning a novel word in their native language, or that novel word's translation), performance was superior on the discrimination task. Interestingly, words trained with novel referents in this task were no different from words trained without any picture at all. The data for the 4-AFC were a little different: a semantic advantage (for known referents) here only emerged for words which were congruent with the participants' native language. Nevertheless, this is clear evidence for some effect of semantics in word learning, at least where one does not have the additional cognitive load of acquiring a novel referent. However, it may be that rather than semantics *per se* being important, more information present at encoding provides more 'anchoring' points for a representation. Essentially, schema-congruent phonological *or* semantics may act as retrieval cues.

In conclusion, the data surrounding semantics are somewhat mixed. Whilst some authors (e.g., Havas et al., 2018; Hawkins et al., 2014; McKay et al., 2008; McNeil & Johnston, 2004) have shown that the provision of semantic information supports lexical representations, this finding has not been extended to measures of lexical engagement. Other authors have shown that the provision of semantic information makes little difference (e.g., Henderson et al., 2013), or have argued that semantic effects emerge later and may delay the emergence of lexical competitions effects (Dumay et al., 2004). This last point seems unlikely, and has not been suggested elsewhere in the literature (e.g., Henderson et al., 2013; McMurray et al., 2017; Tham et al., 2015). Whilst semantic information may not *harm* the emergence of

lexical engagement, clearly, it is also still debatable to what extent (if any) that it supports it.

3.4 Do different learning environments produce qualitative differences in word learning, or affect its time course?

In previous sections, data were presented showing that exposure was an important factor in bringing about lexical competition (Gaskell & Dumay, 2003), and that semantic training may (Havas et al., 2018; Hawkins et al., 2014), or may not (Davis & Gaskell, 2009; Dumay et al., 2004; Henderson et al., 2013), promote lexical engagement emerging. Throughout the chapter, very recent research has been referred to, for discussion later in the thesis (e.g., McMurray et al., 2017; Palma & Titone, 2020). This body of work shows that consolidation – whilst it still may be taking place – is a sufficient but not necessary condition of lexical engagement, as effects may be detectable in the period immediately following learning, providing one uses an appropriate training and testing regime[2]. To preface these chapters, it is worth briefly considering the role that training may play in forming truly lexical representations.

Many adult word learning studies use many repetitions and large numbers of words. Even so, as the literature review above has shown, there is quite some variability in the findings. This in and of itself would suggest that training is not that important, as for example, explicit recognition, e.g., on a 2-AFC task, is usually good, despite different training regimes. The variation across experimental findings may be more likely to be driven by the procedures other than training (e.g., measurement tasks).

However, both research in adults and children have shown that minimal training may be sufficient for a novel word to establish an accessible representation. 'Fast mapping' is an experimental procedure, thought to simulate the early learning environment of children. Experimentally, it is a useful technique as it allows for implicit exposure to a novel word as with phoneme monitoring, but also allows the inclusion of semantics in the study (though cf., Hawkins & Rastle, 2016). Fast mapping has produced interesting findings in adult word learning, and it has been suggested that the procedure allows information to be more rapidly incorporated into neural memory networks, avoiding the need for consolidation and promoting immediate lexicalisation of novel words (Atir-Sharon, Gilboa, Hazan, Koilis & Manevitz, 2015; Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017; Himmer et al., 2017; Merhav et al., 2014; Merhav, Karni & Gilboa, 2015; Sharon et al., 2011; Zaiser, Meyer & Bader, 2019b).

In the canonical fast mapping study, child participants were asked to fetch an experimenter a 'chromium' (olive green) tray. However, the experimenter explicitly contrasted the 'chromium' tray with a red one, by saying to the child "You see those two trays over there? Bring me the chromium one. Not the red one, the chromium

[2]Indeed, in some instances, it is neither a sufficient *nor* necessary condition, as lexical engagement takes more than 24 hours to emerge (see Brown et al., 2012; Coutanche & Thompson-Schill, 2014; Gaskell & Dumay, 2003; Hawkins & Rastle, 2016; Himmer, Müller, Gais & Schönauer, 2017; Walker et al., 2019)

one" (Carey & Bartlett, 1978, p. 18). The characteristic criteria of a fast mapping study are: the presence of one or more familiar referents, a second and novel referent, and a novel word (Carey & Bartlett, 1978; Markson & Bloom, 1997; Halberda, 2006; though see Sharon et al., 2011). Participants then engage in 'referent selection' (i.e., the pairing of a referent and the novel word).

The experimental work for this thesis will begin with fast mapping, as a way of looking at both semantics and faster lexicalisation. Part II of this thesis contains three chapters. Chapter 4 will continue to look at the fast mapping literature, whilst Chapters 5 and 6 (pp. 55 and 65) gauge the paradigm's ability to produce immediate lexical engagement.

Part II

Lexical engagement in 'fast mapped' novel words: does 'fast mapping' lead to immediate lexical engagement?

CHAPTER

FOUR

REVIEW OF THE FAST MAPPING LITERATURE

This chapter will be split into two sections. The first will examine fast mapping (FM) in children. In children, FM is largely concerned with establishing what constrains the word learning 'system' (as it is), and how children acquire words, as they do so at a prodigious rate (e.g., Nation & Waring, 1997). The second will then examine both the behavioural and neuroscientific findings from adult participants, and how this child word learning paradigm has been picked up and applied by memory researchers. In particular, FM learning conditions have been argued to promote immediate lexicalisation of novel words following learning, and suggests a possible route by which the provision of a certain type of semantic information might actually enhance lexical engagement.

4.1 Fast mapping in children

FM was originally conceived as a way of investigating naturalistic word learning in children, whilst also precisely controlling the number of exposures to a word that a child received in an experimental setting (Carey & Bartlett, 1978). Researchers were then able to distinguish between two stages of word learning – a 'fast mapping', initial stage, whereby a phonological code was mapped to a referent, and an 'extended/slow mapping' stage – perhaps equivalent to consolidation in the adult literature – whereby children established more fully-formed representations (Carey & Bartlett, 1978; Carey, 2010; Swingley, 2010). However, unlike in the adult literature, the question of systems consolidation is not prominent, and these data give only limited insight into lexical engagement (e.g., Carey, 2010; O'Connor & Riggs, 2019; Swingley, 2010).

However, the adult literature looking at FM use as their rationale purported findings in the developmental literature (O'Connor & Riggs, 2019), and it is therefore instructive to selectively review some developmental work. Furthermore, questions addressed in the developmental literature are reflected in the work with adults. Two of these debates in particular stand out. The first concerns the role of a competing familiar object in FM trials. Developmental researchers have variously suggested that it is highly supportive of word learning (e.g., Zosh, Brinster & Halberda, 2013), and that increased numbers of objects competing for an infant's attention make word

learning more difficult (e.g., Horst, Scott & Pollard, 2010). This will be discussed in the first subsection (4.1.1, p. 42).

The second subsection (4.1.2, p. 43) will discuss arguments that FM represents no distinct process whatsoever (Dysart, Mather & Riggs, 2016; Kaminski, Call & Fischer, 2004; O'Connor, Lindsay, Mather & Riggs, 2019; O'Connor & Riggs, 2019), and on the contrary, that it is interesting as a particular application of reasoning abilities to support learning (Halberda, 2006).

It should be noted that the questions dealt with in these two subsections are important to distinguish, as even if FM is a general learning ability with no specific mechanism, it may be a feature of the word learning system that competitor objects and/or referent selection support word acquisition.

4.1.1 The role of the competitor in child fast mapping

Horst et al. (2010) have argued strongly that the presentation of multiple competitors at test do not support better retention at test in children. Indeed, they reported that further competitors seemed to actively harm a child's ability to recall the relevant word. Unlike in the adult data, where participants are tested for evidence of lexical competition on large numbers of words trained many times (see Chapter 3, p. 21), Horst and colleagues took the more developmentally-appropriate strategy of training a smaller set of words, and testing children's ability to correctly identify the relevant object (similar to a single non-computerised, X-alternative forced choice trial). Thirty six children aged approximately two and a half years old were divided into three groups. All three groups were trained by being asked to select a novel referent in response to a novel word (e.g., "Where is the dax?"), but differed on the number of familiar competitors present (two, three, or four). The children's response times (RT) on these trials were measured, and found to be significantly faster when only two objects were present, suggesting that the time to select a referent increased by approximately half a second per object present on a trial. Furthermore, this confirmed that even in a condition with multiple objects, participants were not paying less attention (at least, as indicated by RT), as RT scaled with competitor numbers. Accuracy was also uniformly good across all training conditions. All children performed four trials with a novel object and therefore learnt four object-label pairings. Despite this equality of attentiveness across trials, after a delay of five minutes, children were only able to select a target object from the three novel foils above chance where it had occurred in the context of two familiar competitors during referent selection. Children in the other two groups were unable to correctly identify a novel object at test above chance (i.e., where it had occurred in the context of three or four competitors).

To further cement this finding, it was considered a possibility that children may be retaining at least one object (perhaps from earlier in the referent selection trial sequence) – the ability to recall an object at test was therefore analysed as a function of when a child had learnt it (i.e., because objects learnt first or last may have been more salient). An effect was found with this analysis – out of the four items tested, children only tended to recall two of them. However, this finding was again only for objects learnt against two familiar competitors – no such finding was found for

objects learnt against more competitors. Horst et al. (2010) therefore concluded that the presence of further competitors harmed the ability to match a novel object to a heard novel label, and that this was not the result of when a child had learnt the object, or lack of attentiveness, or children being unable to perform referent selection equally well across training conditions.

Although there are differences between the studies, comparison may be made between Horst and colleagues work and that of Zosh et al. (2013), who found that the presence of a single competitor, relative to no competitor at all, supported word learning. This data was collected in slightly older (M_{age} = 38 months) children and with no retention interval, however – testing was performed immediately after training. Testing was the same format, but recognition was only required for a single object in a 4-AFC (Experiment 1, as in Horst et al., 2010) and a 3-AFC (Experiment 2). Training took place by means of either 'direct instruction' ("Point at the dax", only a 'dax' present), or by FM (which the authors call inference, "Can you point at the dax?", where the 'dax' is beside a banana). All inference trials had only a single familiar competitor, so the comparison with Horst et al.'s work is not perfect (as it may be that a single competitor aids, but multiple competitors hinder recognition). Nevertheless, Zosh and colleagues found in two experiments that for five out of 6 objects learnt (although only a single object per child was tested) above chance recognition in the inference condition only. By contrast, the direct instruction condition found chance responding for five out of the six objects. Further analyses suggested that this was not a function of the trial number on which an object had been learnt (even if only a single object had been tested). The authors attributed their findings to three explanations, which they believed to work together: (1) increased interest and engagement by the children in the more difficult inference learning trials; (2) greater depth of processing on these trials also (cf., Craik & Tulving, 1975), and; (3) the support from multiple retrieval cues (e.g., the child remembered seeing the banana with the 'dax', and thus with the relative ease of remembering the (known object) banana, it was easier to remember the (novel object) 'dax'). This final argument is the most interesting, and applicable to the adult findings, as it suggests that semantics support word learning. It should be underlined however that not having a retention interval is a major flaw of this work. It has been argued that a characteristic of fast mapping traces is their susceptibility to interference (in adults; Gilboa, 2019), and data has been provided showing rapid decay of fast mapping memory traces (in both children and adults; Vlach & Sandhofer, 2012).

However, whilst these papers are not entirely mutually exclusive, it seems from the child literature at least that it is difficult to conclude what the role of a competitor might be. This question will be further revisited in the discussion of the adult data (Section 4.2, p. 44).

4.1.2 The distinctiveness of fast mapping in children

A problem with discussing FM is that there is no consensus on precisely what FM is (e.g., contrast Carey & Bartlett, 1978; Sharon et al., 2011), and whether it is distinctive. Two opposing views are summarised in papers by Dysart et al. (2016),

who argues that FM is driven by novelty and extends outside of word learning (cf., Markson & Bloom, 1997) – and Halberda (2006), who argues that FM is the application of reasoning capabilities and a 'process of elimination' logic (disjunctive syllogism) to word learning. Further to this, other authors have argued also in favour of FM being the generic application of novelty matching, as the ability also may be shared with dogs (Kaminski et al., 2004), and operates under the same parameters as any other human faculty involving memory (Vlach & Sandhofer, 2012). The lack of a clear definition in the developmental FM literature may be the cause of some of the misunderstandings, or failures to replicate, in the adult memory literature (cf., O'Connor & Riggs, 2019).

In three experiments, Halberda (2006) first argued that adults engage in disjunctive syllogism (Experiments 1 and 2), and in a final experiment, that the pattern of responding for children between three and four years of age approximated the pattern observed in adults. Children were asked to look and point at one of two screens, which either displayed a novel object (e.g., a 'dax') or a familiar object (e.g., a brush). Children, much like adults, were found to fixate on an object which matched perceived input (e.g., the /d/ in 'dax' does not match the expected /b/ for 'brush', and so shortly after word onset, participants orientated their attention towards the correct object). However, whereas on known trials, the fixations for the target object then persisted, when the target object was novel, after *offset* (e.g., the /s/ of 'dax'), participants looked back to the contrasting familiar object, before returning to the novel target. From this 'double checking' behaviour, Halberda concluded that participants were actively and explicitly rejecting the familiar object. This was confirmed in the adults, who demonstrated meta-cognitive awareness of their behaviour. In doing so, the author strongly rejects accounts of infant word learning that put novelty matching at their centre (e.g., 'N3C' – Novel Name-Nameless Category; Golinkoff, Hirsh-Pasek, Bailey & Wenger, 1992; Mervis & Bertrand, 1994).

However, the appeal of such accounts based on novelty matching is that they do not rely on complicated computations, and fit more widely with accounts favouring simpler, less-specialised processes (e.g., Kaminski et al., 2004; Markson & Bloom, 1997). Recently, Dysart et al. (2016) has shown that children are also able to 'fast map' actions. Pre-exposing some objects (but leaving them un-named/un-actioned), at test the researchers presented again these pre-exposed objects against objects never before seen – super-novel objects. The super-novel objects were reliably chosen in preference to the pre-exposed novel objects as referents for a novel word or a novel action (suggesting shared processing between words and actions in referent selection/FM behaviour; cf., Arbib, 2005). This replicated and extended data from Riggs, Mather, Hyde and Simpson (2015) likewise showing parallels between action-object and word-object mappings in young children.

4.2 Fast mapping in adults

Much of the adult data is concerned with two questions, and will hence be dealt with in two subsections. The first of these addresses questions around whether FM is neurally distinct from other forms of learning, insofar as it is a hippocampally-

independent process (e.g., Coutanche, 2019). Several papers have addressed this question by looking at patients (Korenic et al., 2016; Merhav et al., 2014; Sakhon, Edwards, Luongo, Murphy & Edgin, 2018; Sharon et al., 2011; Warren & Duff, 2014; Warren, Tranel & Duff, 2016). Others still have looked for changes in BOLD (blood oxygen level dependent) signal across the brain, a measure of neural activity in healthy adults, in order to try to localise FM to distinct brain regions (Atir-Sharon et al., 2015; Merhav et al., 2015). As none of these papers are interested in language, *per se*, but words as arbitrary associations between diverse inputs (e.g., Gaskell & Marslen-Wilson, 1997; Warren & Duff, 2014; Warren et al., 2016), these papers do not examine lexical engagement. Lexical configuration (usually in the form of an X-AFC test) is sufficient to tap recognition and therefore provided the data the experimenters wanted. However, although this body of work does not directly speak to the topic of this part of the thesis, it is still important as it provides the arguments and narratives that other authors have probed behaviourally in papers which do feature lexical engagement measures (Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017; Zaiser et al., 2019b).

4.2.1 Neuroscientific findings and fast mapping in the memory literature

A central idea in the FM literature is that semantic schema – structures of knowledge around a topic – may provide 'another way in' to the word learning system (cf., Koutstaal, 2019), particularly for impaired patients. These patients typically have some form of brain damage (e.g., from traumatic brain injury, dementia, or neurosurgery) and any procedure which allows them to overcome their impairment is of course to be welcomed. Tse et al. (2007) provided the groundwork for this. Studying rats, and lesioning their brains, the researchers showed that rats were better able to learn to associate particular flavours with particular locations when these associations had been learnt as part of a schema. This was contrasted with a situation where the rats were forced to learn the associations as arbitrary and isolated 'facts', where learning was poorer. This final manipulation confirmed it was not simply the case that rats trained on the ability to learn the associations became 'expert' (superior) learners, but that there was some effect of schema (for similar work in humans showing an advantage of learning schema-congruent words, see Havas et al., 2018). Critically, rats with hippocampal lesions were still able to use their schema to support learning, as the schema were thought to be stored as distributed cortical representations unaffected by hippocampal lesioning.

Sharon et al. (2011) were the first researchers to apply this finding to memory research. They found that amnesiacs who learnt through FM were able to acquire and recognise arbitrary word-object associations above chance. However, they were not able to perform so in an 'explicit encoding' (EE) condition, where participants were simply told to remember the pairing. This was not the case for neuro-typical controls, who could learn through either method (though for whom hippocampally-mediated EE led to better recognition). This pattern of data was present at testing points ten minutes and one week after training.

It should be noted that authors that are part of this research team (see also Atir-Sharon et al., 2015; Gilboa, 2019; Merhav et al., 2014, 2015) have a conception

of FM not reflected in the developmental literature (e.g., Carey & Bartlett, 1978; Markson & Bloom, 1997; Halberda, 2006). For example, they state the FM task must feature a question (e.g., "Is the numbat's tail pointed [sic] up?", Sharon et al., 2011, p. 1147, Fig. 1a), something not found in the canonical work ("Bring me the chromium one. Not the red one, the chromium one.", Carey & Bartlett, 1978, p. 18; 'explicit disjunctive syllogism', Halberda, 2006). Some authors (e.g., Merhav et al., 2014) have argued that without a question the difference between FM and EE is not realised (despite developmental implementations, and adult behavioural data showing that the question alone does not make a difference; Coutanche & Thompson-Schill, 2014). Nevertheless, this paper provided the first putative evidence that FM could produce unexpected results in adults, and was evidence against the complementary learning systems model (CLSM) account (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). As the patients in Sharon et al. (2011) had hippocampal damage, it was thought that the schema activated by the familiar competitor allowed the new word-object pairing to be learnt without consolidation/hippocampal involvement. Under this model, the activated schema would allow for rapid integration of the cortical representation (produced at encoding) into the lexicon. Early evidence for this was that two of the patients in Sharon et al.'s work had co-morbid anterior temporal lobe (ATL) damage, alongside extensive hippocampal damage. However, these patients failed to show learning through FM, suggesting that it relied on different brain areas (cf., Atir-Sharon et al., 2015; Merhav et al., 2015).

The ATL has been implicated in semantic processing as an amodal hub for the integration of diverse aspects of a representation – an idea picked up and tested further in an FM context by Merhav et al. (2014, 2015) and Atir-Sharon et al. (2015). In brief, these papers have found neuropsychological differences between FM and EE which support this idea. However, there has not been uniform support of this work. Warren and Duff (2014) failed to replicate the findings of Sharon et al. (2011), with their amnesiacs unable to learn from either FM or EE, but healthy controls able to learn from both training types. Moreover, Warren and Duff found that amnesiacs also performed poorly during FM training, and speculate this is due to the hippocampus acting as a binder of information (e.g., Teyler & DiScenna, 1986; McClelland et al., 1995). Amnesiacs, they argued, struggle to relate the information present during training (i.e., novel word and object) to complete the task successfully, due to their impairments. Further study with patients with temporal lobectomy confirmed the lack of an FM advantage, although this time, patients' ability to complete training was intact (Warren et al., 2016). Warren and Duff (2019) would later further emphasise their results by arguing that all word learning not involving the hippocampus is "slow and sparse, irrespective of methodology" (p. 1). By contrast, Gilboa (2019), a researcher involved with the earlier FM work (Atir-Sharon et al., 2015; Merhav et al., 2014, 2015; Sharon et al., 2011), would cease to claim that FM mapping produces *better* memory, and instead argue that FM produces *different* memory; a central characteristic of which is fragility and susceptibility to interference.

4.2.2 Behavioural findings: fast mapping and rapid lexicalisation

In contrast to the neuroscientific literature, the behavioural literature features both configurational and engagement measures. In the engagement literature in particular, FM is seen by its advocates as a way of generating representations which may be integrated more rapidly with existing knowledge (Coutanche & Thompson-Schill, 2014, 2015; Coutanche & Koch, 2017; Zaiser et al., 2019b). This work has however proven difficult to replicate (Gaskell & Lindsay, 2019)[1]. The behavioural literature supports the neuroscientific literature by looking at various components of the FM task (cf., Cooper, Greve & Henson, 2019b).

Studies reporting a 'fast mapping effect'

Studies with lexical configuration measures. In support of the uniqueness of FM, seven papers are of note. Firstly, in addition to his work with infants, in the same work Halberda (2006, Experiments 1 & 2) showed a similar pattern of data with adult participants. Arguing that referent selection in FM draws on disjunctive syllogism (DS, 'not A, therefore B' logic), in his first two experiments, Halberda contrasted implicit ("The winner is the 'dax'") and explicit DS ("The winner is *not* the iron"). Halberda argued that participants were logically working through the rejection of the familiar referent during referent selection, as in both conditions participants performed a 'double check' (looking away from the novel object and then returning to it).

Secondly, Havas et al. (2018) reported data showing that schema may support lexical representations (cf., Tse et al., 2007). Testing Spanish native speakers, participants were taught either: a novel word conforming to phonological rules of Spanish, or a novel word conforming to the rules of Hungarian. As Hungarian has phonemes not found in Spanish, it was expected that the Spanish speakers would show weaker knowledge for these novel words, as they were less able to encode and store them. Additionally, the words were trained with either a familiar referent (e.g., a cat) or a novel one (e.g., an unusual artefact), to create a 2 × 2 design. The researchers found that where a schema was present (*either* in words with familiar phonology *or* with a familiar referent), performance was improved on a 4-AFC measure. Finally, on a semantic priming measure, no effect was found, but for words with a known phonology (although not before participants had slept, consistent with consolidation accounts; Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). Whilst this paper did not study FM, it supports a key tenet of the argument – namely, that schema support word learning, and may even accelerate engagement. Indeed, this idea has more recently been picked up in updates to CLSM (McClelland, 2013; Kumaran et al., 2016; McClelland et al., 2020). The role for schema is further supported by data from Zhang, Popov, Koch, Calloway and Coutanche (2018), who found that integration of learnt paired associates A — B and B — C into A — B — C was facilitated by schema consistency (between a person and a place, e.g., teacher — classroom, but not baker — mountain).

[1] Gaskell and Lindsay make an oblique reference to unpublished work failing to replicate the work of Coutanche and Thompson-Schill and Coutanche and Koch. Zaiser et al. is also an unpublished preprint manuscript. It should be emphasised that no failures to replicate have yet been published.

Inspired by the neuroscientific data, Himmer et al. (2017) set out to look at the effect of sleep-mediated consolidation on FM memory traces. The 3-AFC task was used to probe declarative memory. The authors found that only memory traces formed by EE saw overnight improvements, consistent with the neuroscientific arguments that fast-mapped traces are stored cortically, and thus do not need to be consolidated into cortical memory. Likewise, as reported in the neuroscientific literature, the cost of this immediate integration was weaker declarative memory (cf., Atir-Sharon et al., 2015; Merhav et al., 2015; Sharon et al., 2011). However, it may simple be that traces learnt by FM are too weak to see a consolidation benefit with a single night of sleep (Walker et al., 2019).

Studies with lexical engagement measures. Testing across two days, and comparing EE to FM in undergraduates, Coutanche and Thompson-Schill (2014) found that FM produced a lexical competition effect *before* sleep, similar to that observed by Bowers et al. (2005; indeed, using their word lists) *after* sleep. This could be taken as evidence of a 'FM advantage'. However, the authors also found that FM produces weaker declarative knowledge, with participants performing better on a 3-AFC task for items learnt by EE. The authors' implementation of FM was similar to that seen in the neuroscientific literature (e.g., Sharon et al., 2011), with a familiar referent contrasted with a novel referent, in the presence of a carrier question introducing a novel word (e.g., "Are the antennae of the 'torato' pointing up?"). In contrast to previous work, however, the authors did not use the usual fruits, flowers, animal item set (seen in e.g., Atir-Sharon et al., 2015; Cooper et al., 2019c, 2019b; Greve, Cooper & Henson, 2014; Merhav et al., 2014, 2015; Sharon et al., 2011; C. N. Smith, Urgolites, Hopkins & Squire, 2014; Warren & Duff, 2014; Warren et al., 2016).

Instead, they used their own pictures of novel animals. This allowed Coutanche and Thompson-Schill to more deliberately emphasise during learning that the novel and familiar objects were of the same taxonomic class (i.e., not just two animals – which may be quite different, e.g., a bird and a reptile, but specifically, two insects). The authors were therefore able to make a stronger argument about the importance of schema activation (drawing on data from Tse et al., 2007). Interestingly however, despite their effect supposedly relying on some degree of semantic processing (through the schema activation), no evidence of semantic priming was found until after participants had slept. Nevertheless, it was the case that this was only seen in the FM condition – EE failed to bring about any semantic priming at all over the two days of testing, so some 'FM effect' was still suggested. Despite this, though, the findings remained confused, and unaccounted for – the authors did not address, or explain, why it may be the case that *semantic* processing allows only fast *phonological* integration[2]. Particularly from a DCM perspective, it is not clear how or

[2]It could be the case that semantic connexions generally are slower to emerge; cf., Dumay et al. (2004). However, this is not conclusively the case – Tham et al. (2015) found evidence of immediate semantic integration, albeit with different measures. Coutanche and Thompson-Schill's findings may be some artefact of task – and for example, semantic integration may be boosted also – but this is speculative. The point remains: it is *not* addressed in their paper why a purportedly *semantic* effect would induce faster *phonological* abstraction and generalisation.

why semantic overlap would facilitate integration of only the phonological aspect of the novel representation.

In a second experiment, Coutanche and Thompson-Schill introduced a third (and hereto novel) condition, 'implicit encoding'. This took the question from the FM condition, and participants were presented with this in the presence of a single object, as seen in the EE condition. However, when comparing the three conditions, they found that only the FM condition was clearly distinguishable, with explicit and implicit encoding sitting together on lexical configuration and engagement measures. They therefore concluded that the role of the competitor was crucial, linking to the developmental data (e.g., Zosh et al., 2013).

Coutanche and Koch (2017) later followed up this work. Believing the extent to which a participant relied on semantic memory to be subject to individual differences, they compared participants who relied more, or less, on semantic memory. Participants not relying on semantic memory were deemed to be drawing more on an episodic system centred on the hippocampus. Categorisation of participants was decided by their scores on the Survey of Autobiographical Memory (SAM; Palombo, Williams, Abdi & Levine, 2013). Coutanche and Koch also manipulated competitor typicality. They predicted that participants who were using semantic memory, would learn better when the competitor was atypical (e.g., penguin, rooster, ostrich, and chicken were atypical birds, whereas sparrow, blackbird, robin, and dove were typical, as found in a previous pilot study). Their data showed that participants in the bottom half of the SAM score distribution (thought to be drawing on episodic memory) showed no competition effect, for either typical or atypical competitors. By contrast, in the top half of the SAM score distribution, words learnt by FM on trials with typical competitors induced lexical facilitation, and words learnt on atypical competitor trials showed evidence of lexical competition. As in Coutanche and Thompson-Schill (2014), these data were found before sleep – no testing was performed after sleep. The control conditions, of explicit and implicit encoding, again showed no effect. No effect was found for semantic priming, again, as in Coutanche and Thompson-Schill's earlier work.

Studies reporting no 'fast mapping effect'

Seven studies report behavioural data finding no evidence of better learning under FM conditions. Four of these are with specific participant populations, questioning the veracity of the neuropsychological and neuroscientific data in particular, and three further studies report findings with neurotypical adults, reflecting on the purported cognitive mechanisms supporting FM in studies such as those by Coutanche and colleagues (2014; 2017). It is worth noting, however, that no studies have been published showing a failure to replicate the lexical engagement effects[3].

Studies with particular groups. A wide range of particular interest groups have failed to find benefits to learning by FM. Of particular interest is the direct

[3]Gaskell and Lindsay (2019) makes reference to two unpublished failures to replicate; Cooper, Greve and Henson (2019a) also have a pre-registered, but unpublished, failure to replicate. This work was only performed subsequent to Experiments 1 and 2.

replication of Sharon et al. (2011) by C. N. Smith et al. (2014). Controls and patients with brain damage were compared ten minutes after learning by either EE or FM, and then again after a week. Patients were typically around chance performance, and consistently outperformed by controls. Controls performed significantly better under EE conditions. This data is consistent with the side of the argument articulated most forcefully by Warren and various colleagues (2014; 2016; 2019) against any benefit for learning by FM, to patients or controls. Studies in other groups of patients (those with schizophrenia; Korenic et al., 2016; and down syndrome; Sakhon et al., 2018), who are likewise thought to have impaired brain and/or learning functions, have shown similar failures to find an FM effect. Lastly, taking the view that in normal ageing brain volume decreases and is (possibly) associated with a similar decline in cognitive function, Greve et al. (2014) compared old and young participants. The older participants had smaller hippocampi ($M = 3.92\text{cm}^2$, $SD = 0.49\text{cm}^2$) relative to younger participants ($M = 4.44\text{cm}^2$, $SD = 0.41\text{cm}^2$), in addition to being older ($M_{old} = 66.0$ years, $SD_{old} = 6.3$ years; $M_{young} = 26.9$ years, $SD_{young} = 7.4$ years). Performance on a 3-AFC task ten minutes, and a week, after learning showed consistently better performance in the EE condition, for both groups. Likewise, hippocampal volume was found to predict both FM and EE, to the same degree – a relationship that should not exist if the argument that FM leads to immediate cortical integration is to be accepted. Casting doubt on the neuroscientific data (e.g., Atir-Sharon et al., 2015; Merhav et al., 2015), which implicated the anterior temporal lobe (ATL) in FM, ATL volume was not found to predict either FM or EE performance, in either participant group.

Studies with neurotypical adults. Three studies are of interest with neurotypical adults. The first, from Vlach and Sandhofer (2012), used an implementation of the FM paradigm more applicable to the developmental literature, and featured the introduction of only a single word 'koba'. Participants (three year-olds, and adults) were asked to play a game which involved the measurement of 6 novel objects. Five of the novel objects were referred to without a specific label ('this', 'it', 'toy'), and a single object was labelled as 'koba', to participants unaware that they were in fact partaking in a word learning experiment. At time points immediately, one week, and one month post-test, both adults and children experienced forgetting (inability to correctly select the referent 'koba' object and reject the five novel distractors – akin to a single 6-AFC trial). Crucially, forgetting was in a familiar exponential decay curve (e.g., Ebbinghaus, 1913), suggesting that although participants might show good retention immediately after testing, it was not the case that this promoted superior or different learning in the long term. Interestingly, after a month, children and adults' performance was equivalent, and at less than 20% accuracy.

The second study is elegant, insofar as it decomposes elements of the FM task found in the adult (e.g., Sharon et al., 2011), but not developmental (e.g., Carey & Bartlett, 1978; Markson & Bloom, 1997; Halberda, 2006) literature – inference, and competitor. Across four experiments, combining various participants and conditions, Cooper et al. (2019b) compared:

- Sharon et al.'s (2011) FM paradigm, as seen elsewhere in the adult neuroscience

literature, with a schema-congruent, familiar competitor, and a question requiring participants to make an inference to map novel word to novel object (e.g., "Is the numbat's tail pointed [sic] up?", two mammals shown, one of which is novel);

- An FM condition without the competitor (a single novel referent shown, and a question, e.g., "Does the 'loris' have large ears?");
- An FM condition without the inferential question (but both familiar competitor and novel referent shown) – participants instead asked a question like "Is the 'kobus' you see on the right familiar?";
- A condition with neither competitor nor inferential question – participants saw a single novel referent, and were asked a question like "Is the 'culogo' you see on the right?";
- A standard EE condition (a single novel referent, shown with the instructions to remember it, e.g., "Remember the 'tarsier'").

The decomposition of the FM task showed no changes in 3-AFC performance. In all cases, Bayesian statistics preferred the null hypothesis, although in some experiments, performance actually improved when task demands (e.g., due to the removal of question or competitor) were diminished. The authors suggested that with an easier task, more resources could be devoted to learning itself, resulting in a stronger/better-encoded memory trace.

The last, and arguably most important study in this review was drawn together by Cooper et al. (2019c). In response to various concerns about the veracity and robustness of the FM findings, the authors undertook a review of the experimental evidence similar to that framed above. Additionally, however, the paper was important for the field as commentaries were invited from a wide range of laboratories and research groups, offering a range of perspectives (adult memory, psycholinguistic, developmental, neuroscientific, etc.). Their positions with respect to the conclusions of Cooper et al. (2019c) are summarised in Table 4.1 (p. 53), which were then replied to in the researchers response to those commentaries (Cooper et al., 2019d). In their review of the literature, Cooper et al. (2019c, p. 12) were explicit:

> "In healthy adults (with an intact hippocampus), there is currently no evidence of faster or better integration of new information under FM than EE in tests of explicit memory [i.e., lexical configuration]. Additionally, the limited evidence that exists for an FM advantage in tests of implicit memory raises several additional theoretical puzzles, and deserves further replication. The question of whether [FM] occurs in adults thus remains unresolved, much like the question of whether [FM] is really a distinct form of learning in the developmental literature from where the concept originated."

This point was further reinforced in their reply to the commentaries (2019d, p. 240):

> "In conclusion, we stand by our original claim that the evidence for [FM], at least in adults within the [paradigm] introduced by Sharon et al. (2011), is not convincing, and we are comforted that most of the commentators seem to agree with this."

4.2.3 Overview of Experiments 1 and 2

Chapters 5 and 6 (pp. 55 and 65) present the first experimental work of this thesis, drawing upon the work reviewed above, and in Chapter 3. The experimental work contained in these chapters was conducted beginning in late 2016, with testing beginning in the summer of 2017, at a time when much of the more critical literature (e.g., Cooper et al., 2019a, 2019b, 2019c, 2019d) was not quite so solidified, and largely, had not been published. Furthermore, as of September 2021, it remains the case that no failure to replicate the lexical engagement effects of Coutanche and Thompson-Schill (2014) and Coutanche and Koch (2017) has been published. Experiments 1 and 2 sought to further explore these papers, and fast mapping, with respect to lexical engagement.

Table 4.1 Summary of commentaries to Cooper et al. (2019a)

Citation	Position	Summary of comments
Coutanche (2019)	Reject	Cooper et al.'s criticisms are misrepresentations; more nuance needed
Elward et al. (2019)	Accept	Developmental amnesiacs show no FM benefit
Gaskell and Lindsay (2019)	Accept	FM is not special as pre-sleep engagement is observed elsewhere
Gernsbacher and Morson (2019)	Accept	FM is only a laboratory task
Gilboa (2019)	Reject	FM produces different, not superior, memory
Koutstaal (2019)	Neutral	Focus on FM may have led to study of other ways of supporting learning being neglected
Mak (2019)	Reject	Presents a computational account of how a competitor may support learning
O'Connor et al. (2019)	Accept	FM has been misunderstood by adult memory researchers
Warren and Duff (2019)	Accept	Word learning requires a functional hippocampus
Zaiser et al. (2019a)	Reject	Differences across studies may be accounted for by un-elucidated moderating factors

Note. The position of the authors is evaluated from their commentaries with respect to Cooper et al.'s (2019c) arguments. Fast mapping abbreviated to FM.

CHAPTER

FIVE

EXPERIMENT 1
SCHEMA ACTIVATION IN FAST MAPPING

5.1 Introduction and rationale

Previous research has shown no conclusive findings with respect to the effect the provision of semantic information during word learning, as discussed in the previous chapters (cf., Hawkins et al., 2014; Henderson et al., 2013). The evidence is quite mixed – although it does seem possible to eliminate accounts that suggest that semantic information slows the emergence of lexical engagement (Davis & Gaskell, 2009). The fast mapping (FM) paradigm allowed further exploration of the relationship between semantics and lexical engagement. The paradigm was particularly suitable as there was a body of literature suggesting that semantic information supported lexical engagement (Coutanche & Thompson-Schill, 2014, 2015; Coutanche & Koch, 2017), bringing it about faster than predicted by models of word learning (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). Moreover, its ecological validity was also appealing: FM simulates the natural word learning environment (e.g., Carey & Bartlett, 1978), in contrast to more abstracted and impoverished learning in other studies (e.g., Gaskell & Dumay, 2003).

Recent evidence from the FM paradigm suggested that learning in this manner could accelerate the time course for lexicalisation (Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017). The mechanism for this apparent finding is a schema, activated and shared across old and new information (e.g., integrating the name of a new insect into the lexical network is easier and faster if it is placed next to a known insect; see also Havas et al., 2018; McClelland, 2013; Tse et al., 2007). Specifically, in their second experiment, Coutanche and Thompson-Schill (2014) argued that schema activation in FM was driven by the familiar competitor object during training. Later work would suggest that this was particularly true for *atypical* objects (Coutanche & Koch, 2017; Coutanche, 2019). The same experiment also suggested that the other aspect of the FM task – the semantic question introducing the novel word (i.e., "Are the antennae of the 'fostil' pointing up?") – did not lead to lexical engagement, unless the competitor was also present. Note that the question participants had to make responses to was designed to draw attention to a feature

shared between the competitor and the target (here, antennae), although it was not always diagnostic of that taxonomic class (as arguably antennae are of insects – other trials made reference to legs, wings, etc.).

It is here that the argument begins to break down, as Coutanche and Thompson-Schill (2014) do not clearly articulate why a shared *semantic* feature might bring about better *lexical* integration, where it is indexed by lexical competition. In the case of a known referent (e.g., GRASSHOPPER) boosting/activating a schema (presumably, INSECTS), allowing the rapid integration of a novel word (e.g., 'fostil', and its referent, a giraffe-necked weevil, see Fig. A.1, p. 213), it is not clear how the *lexical* competitor ('fossil') becomes linked, such that the new word competes with it for activation, and slows its recognition in a semantic categorisation task. Moreover, it is not clear why the novel referent – always recognisably from a particular taxonomic class – cannot activate the referent schema (e.g., INSECTS) on its own, particularly when paired with a question which draws attention to the features of the referent. Why should a competitor be required when the referent is recognisably of a class/schema into which the novel object will be integrated?

Leaving aside the question of *what* feature activates the schema, and why *atypical* animals (e.g., 'penguin') would more strongly activate their schema (e.g., BIRDS), when they have less featural overlap, there is also a problem of a missing design cell – only three out of four experimental conditions pairing a semantic question and a competitor have been tested (Coutanche & Thompson-Schill, 2014). These were:

- A competitor and a semantic question (FM);
- No competitor and no semantic question (explicit encoding, EE);
- No competitor and a semantic question (implicit encoding, IE; the new condition in Coutanche and Thompson-Schill's second experiment).

Missing was a condition where there was a competitor, but no semantic question. Only by including this condition and still finding a lexical competition effect could one conclude conclusively that the competitor was central to the effects.

Another oddity of the adult FM literature is that it is precisely this final design cell that is in the developmental literature from which the FM effect is supposedly drawn. Many developmental studies use no question at all (Dysart et al., 2016; Horst et al., 2010; Riggs et al., 2015; Vlach & Sandhofer, 2012), or ask a question that does not make reference to a semantic feature (Halberda, 2006; Zosh et al., 2013). These questions typically require some response that is more spatial than semantic (e.g., 'Where is the...', 'Can you look at the...', 'Can you find the...'). This is contributes to a perceived misrepresentation and overextension of the developmental literature by adult FM researchers (O'Connor & Riggs, 2019).

Experiment 1 extended Coutanche and Thompson-Schill (2014) by completing this cell of the design, and using a task from the developmental literature to do so (e.g., Dysart et al., 2016; Riggs et al., 2015). The central question was what drove schema activation; in the first instance, evidence of lexical competition would be accepted as evidence as schema activation, following the arguments put forward in the literature (Coutanche & Thompson-Schill, 2014, 2015; Coutanche & Koch,

2017; Coutanche, 2019; Havas et al., 2018; McClelland, 2013; Merhav et al., 2015; Sharon et al., 2011; Tse et al., 2007). In Experiment 1, participants were trained by referent selection with Coutanche and Thompson-Schill's stimuli, but were asked "Where is the fostil?" (left/right button press response required). If, as argued, the question does not contribute to schema activation, then this change would not affect lexical integration, and a competition effect would still be detected. By contrast, if the question was central to the effects, against what was claimed by Coutanche and Thompson-Schill on the basis of their second experiment, then no competition effect would be observed.

5.1.1 The present study

The effect of interest in Coutanche and Thompson-Schill (2014) occurred on the first day of testing, and only in the FM condition. Experiment 1 therefore did not collect data from either of their other two conditions, EE or IE. Furthermore, it was not deemed necessary to look for the effect on the second day of testing – Coutanche and Thompson-Schill report that it was maintained for FM in any case. These changes were made as a goal of this experiment was to quickly establish the robustness of the FM effect to alterations in task set up, and pragmatically, it was easier to run a shorter experiment with fewer conditions.

Another small change was also implemented. The experiment followed on from other work at the University of York (unpublished, but referred to in Gaskell & Lindsay, 2019). There, a change had been made to cut the number of items learnt from 16 to 12, given the low number of exposures during training. This change was maintained here also, in a deviation from the original work of Coutanche and Thompson-Schill.

5.2 Methods

5.2.1 Participants

Fifty three participants contributed data (M_{age} = 21.7 years, SD_{age} = 7.40 years). Of these, 10 were male, 43 were monolingual, and 42 were right-handed. All participants were fluent in English. Participants were all tested in a quiet laboratory environment. All were free of any confounding disorders (e.g., sensory, learning or language difficulties), or had corrections to normal (e.g., by wearing eyeglasses).

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, or in exchange for course credits.

5.2.2 Materials and apparatus

Novel referents were 24 little-known animals. A variety of mammals, birds and insects were used, using a set of stimuli received from Coutanche (2014). These had been closely cropped and set against a white background (see Fig. A.1, p. 213). Competitor referents were processed and presented in the same way. All images were in full colour, and participants saw photo-realistic depictions of the animals.

All words used in the experiment had previously been used in published research (Bowers et al., 2005; Coutanche & Thompson-Schill, 2014). All familiar test items were 'hermit words' – that is, words from which no other English orthographic form could be constructed by the addition, substitution, or replacement of a single letter. All novel competitors were constructed by replacing a single letter (e.g., 'walnut' → 'walnot'). There were also a number of filler items. A full list of words used in the experiment can be seen in Table A.1 (p. 213).

For the familiarity test at the end of the experiment, images of the novel referents were printed out in black and white (for cost reasons), and participants filled out this pack on paper.

5.2.3 Design

Critical stimuli were organised into two lists of 12 items, and participants were allocated to a single list. Each list contained novel competitors and familiar words, which either remained 'hermit words', or became 'former hermits'. Hermit words were words for which no novel competitor had been learnt. For example, as a List 1 participant had not learnt the novel competitors on List 2, the familiar words on List 2 were still hermits following training, for that participant. By contrast, for the same List 1 participant, the familiar List 1 words were now 'former hermits', as a competitor had been learnt (see Table A.1, p. 213). All participants saw the same filler items, which were common nouns.

Item order in all tasks with except for the familiarity check was randomised; left and right responses were also appropriately counter-balanced.

5.2.4 Procedure

Procedures used throughout had been adapted from Coutanche and Thompson-Schill (2014). The experiment had two phases: training, followed by testing. Training of word used an adapted FM procedure. Testing consisted of a lexical engagement task (semantic categorisation), a lexical configuration task (three-alternative forced choice; 3-AFC), and a post-test familiarity check. The computerised tasks were scripted in DMDX (Forster & Forster, 2003).

Training

After being allocated to a list, participants began by completing 32 training trials, arranged into two blocks. Each block of 16 trials featured 12 novel word referent selection trials (one per to-be-learnt word, see Table A.1, p. 213). Twelve to-be-learnt items was slightly fewer than Coutanche and Thompson-Schill's 16 items; this change was made due to scepticism over how well participants would perform with so many words to learn and so few exposures (as justified by the literature, cf., Greve et al., 2014; Warren & Duff, 2014; Warren et al., 2016). The remaining four trials per block were familiar catch trials. The purpose of these catch trials was to ensure that participants had to read the question before responding, and were not just seeking the novel object. A potential problem with changing the question at training was that participants could respond entirely correctly by looking for the

novel object, without needing to read the to-be-learnt word printed in the question. However, as on familiar catch trials both objects were known, participants would only know which to select having read the question and looked at both objects.

On novel word referent selection trials, participants saw a novel referent and a familiar referent from the same taxonomic class (e.g., a giraffe-necked weevil, and a grasshopper; for example, see Fig. A.1, p. 213). On familiar catch trials, both objects would be familiar. On both types of training trial, underneath these two objects was a question, "Where is the X?", where X was either a novel or familiar word, according to trial. Participants responded by pressing a key on either the right or the left of the keyboard. The run of 16 trials would play in a random order, before being looped, giving participants two presentations of each novel word-referent pairing. After making a response the object would be held on screen for 6s; if no response was made during this time, the trial was discarded. An analysis of participants' accuracy was performed. Following learning, participants watched a ten-minute video, in order to create a retention interval and to suppress active rehearsal of the learnt words.

Lexical engagement task

Participant lexical engagement was assessed with a semantic categorisation task (Bowers et al., 2005). This required participants to make speeded responses to words, categorising them as being either man-made or natural (half of each across the whole item set). Data were taken from 48 trials (12 hermits, 12 former hermits, 24 fillers). Accuracy and response times (RTs) were examined. Whilst responding, participants saw the word on screen, centralised and printed in large black font on a white background, but no picture. The response cues 'man-made' or 'natural' occurred at the bottom of the screen.

Lexical configuration task

Next was a 3-AFC task, to assess participants' recognition memory. As before, RT and accuracy data were taken. Participants saw three referents which they had learnt on screen, and were presented with a word they had also learnt at the bottom of the screen. Participants had to press one of three buttons to identify the location of the correct referent as being either on the left, in the middle, or on the right of the screen. Each novel referent appeared as a foil for two other objects.

Post-test familiarity check

After all data had been collected, participants filled out the familiarity questionnaire, assessing their pre-test familiarity with the novel referents on a seven point Likert scale (1: 'not at all familiar', 7: 'very familiar'). Referents which had pre-test familiarity, or participants that were familiar with many of the referents, were eliminated from the experimental dataset.

Processing of the data and exclusions

Training. Trials were eliminated from the analysis if participants responded incorrectly, if no response was made within 6s, or if responding was deemed anticipatory (RT < 300ms). This resulted in the elimination of ∼3% of trials. It had been decided before analysing the data that any participants with less than 75% accuracy would be eliminated; however, the minimum accuracy was 87.5% (maximum: 100%). Participants were also considered for elimination by their familiar catch trial accuracy. Responding at chance levels ($\frac{4}{8}$) on these trials was to be taken as evidence of insufficient attention being paid by participants during training. However, all participants responded correctly on at least six trials.

Lexical engagement. The analysis of the data was conducted following a data cleaning procedure used in earlier research (Bowers et al., 2005; Coutanche & Thompson-Schill, 2014). Around 6% trials were excluded in the first instance as they had an incorrect response. Secondly, the data from the familiarity questionnaire were examined, and a further ∼6% individual trials where the participants had said they were not unfamiliar with the novel referent were excluded. Finally, a further ∼9% trials with RTs over 1500ms and under 300ms were also excluded. Subjects were then excluded if they were not unfamiliar with at least half of the novel referents. This excluded eight subjects (∼ 11% of trials). A further eight subjects were eliminated due to having less than 70% of their trials remaining after trial exclusion procedures (∼ 9% of trials). This left a final dataset of 1485 trials across 37 participants – 84% of trials from the remaining participants.

Lexical configuration. Again, trials were eliminated from the analysis if the response was incorrect (trials removed: ∼35%) or anticipatory (RT < 300ms, < 1%). Participants excluded from the lexical engagement analyses were allowed to contribute their lexical configuration, as no participants had been excluded during training.

5.3 Results

All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016).

5.3.1 Training

Twenty six participants were assigned to List 1, and 27 to List 2. Responding was found to be significantly above chance responding (50%; Wilcoxon rank sum test due to non-normality; $U = 2809$, $p < 0.001$, $r = 1.31$[1]).

Participant accuracy across each of the training lists was very similar ($M_1 = 97.5\%$, $SD_1 = 4.76\%$; $M_2 = 97.8\%$, $SD_2 = 4.41\%$), and statistically indistinguishable ($U = 356$, $p = 0.932$, $r = 0.012$). Given equivalent performance in learning the words, List 1 and 2 participants were pooled for all subsequent analyses.

[1] $r = \frac{Z}{\sqrt{N}}$; Rosenthal (1994); this formula will be used throughout this thesis unless otherwise specified.

5.3.2 Lexical engagement

Lexical engagement accuracy data

A summary of the accuracy rates across types of words can be seen in Fig. 5.1a (p. 62). Accuracy rates were highest for filler words ($M_f = 88.7\%$, $SD_f = 6.94\%$), and lowest for hermits ($M_h = 80.8\%$, $SD_h = 15.7\%$). Despite a non-normal distribution of errors, (according to a Shapiro-Wilkes test; $W = 0.815$, $p < 0.001$), the apparent differences were subjected to parametric testing, in line with statistical best practice for samples of this size[2] (Blanca, Alarcón, Arnau, Bono & Bendayan, 2017; Ghasemi & Zahediasl, 2012; Lumley, Diehr, Emerson & Chen, 2002). This showed a significant difference between the three trial types ($F(2, 72) = 3.93$, $p = 0.024$, $\eta_g^2 = 0.070$[3]).

A planned paired-samples t-test was performed between former hermit and hermit words. This showed no difference, suggesting that the effect was instead driven by the filler words ($t(36) = 0.137$, $p = 0.892$, $d = 0.034$[4]). Furthermore, this was weak evidence of a lack of a competition effect[5].

Lexical engagement RT data

A summary of the RT data across trial types can be seen in Fig. 5.1b (p. 62). Responses were fastest when participants saw a filler word ($M = 816$ms, $SD = 105$ms), and slowest when participants saw a former hermit word ($M = 846$ms, $SD = 104$ms). These data were compared by a one way ANOVA, which showed that RT did vary with trial type ($F(2, 72) = 3.57$, $p = 0.033$, $\eta_g^2 = 0.013$). However, a planned comparison showed that this was not driven by a former hermit/hermit word competition effect ($t(36) = 0.819$, $p = 0.418$, $d = 0.095$). This was further evidence against lexicalisation of the novel words.

Lexical engagement supplementary analyses

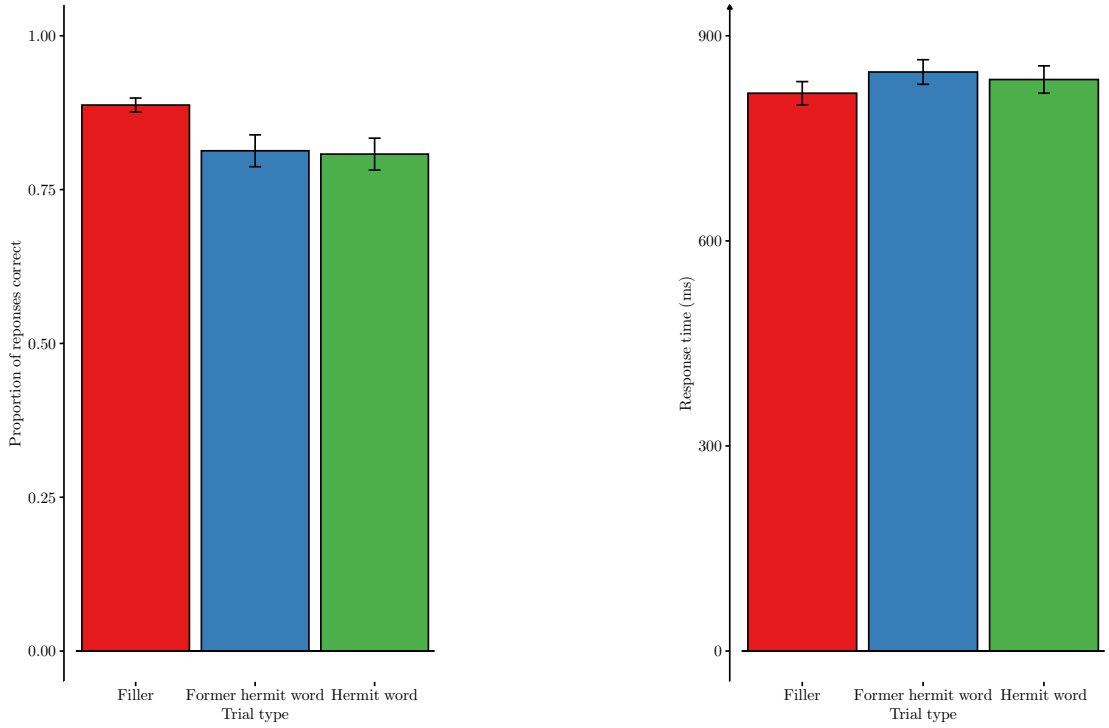
With there being no evidence of lexical competition, it was important to try to eliminate problems with the design as a possible cause, given that Coutanche and Thompson-Schill (2014) had previously shown an effect, which was then further replicated by Coutanche and Koch (2017). It was possible that a drop in the number of items from 16 to 12 here reduced statistical power. To consider this further, the data were examined on a per-participant basis. The 'insufficient power' argument essentially postulates that a true effect in this sample was obscured by random selection of participants, more of whom just so happened not to be good enough learners to detect an effect with this particular item set. This problem can be overcome

[2] Note that this approach will be continued throughout the thesis, as all samples are of sufficient size for parametric testing.

[3] General eta squared, Bakeman (2005)

[4] $d = t\left(\frac{2(1-r)}{n}\right)^{\frac{1}{2}}$; Dunlap, Cortina, Vaslow and Burke (1996), p. 171, Eq. 3. This formula will be used throughout for paired-sample tests unless otherwise specified.

[5] The evidence in accuracy data for competition is mixed, and Coutanche and Thompson-Schill (2014) do not perform an accuracy data analysis. However, Bowers et al. (2005) suggests that a significant difference is evidence of inhibition of the former hermit words by the novel competitors.



(a) Accuracy data

(b) RT data

Figure 5.1: Accuracy and RT data for the lexical engagement task in Experiment 1. Error bars show ±1 *SE*.

by specifically examining participants who showed an RT difference between former hermit and hermit words, which would usually be indicative of lexical engagement. However, to ensure that one is not just selecting one half of a normal distribution around a mean of zero, other indicators of competition must also be looked for. For example, if a particular set of participants shows an RT difference, the size of that difference should also correlate inversely with accuracy on former hermit trials, if the difference is truly due to difficulties with lexical processing on those trials[6].

In the dataset of 37 participants, 16 showed a positive RT difference (former hermit − hermit RT) of, on average, 82ms ($SD = 71$ms). On average, these participants were 81.5% accurate ($SD = 13.6\%$). However, there was no significant relationship between participant former hermit trial accuracy and the size of the RT difference (Pearson's $r = 0.352$, $p = 0.181$). This is further evidence that there was no reliable lexical competition in Experiment 1.

5.3.3 Lexical configuration

Participants completed a 3-AFC task to test their recognition of the novel referents. Overall, 64.5% of novel referents were correctly identified ($SD = 20.2\%$). The average

[6]Note that Coutanche and Thompson-Schill (2014) do not perform a statistical analysis of their accuracy data in either experiment; however, there is a numerical difference. Bowers et al. (2005) record a statistically significant difference in their accuracy data.

RT was 3344ms ($SD = 463$ms). Recognition accuracy was significantly above chance ($\frac{1}{3}$; $t(52) = 11.2$, $p < 0.001$, $d = 1.54$[7]).

5.4 Discussion

In Experiment 1, the possibility of immediate lexical engagement (by lexical competition) under FM learning conditions was explored. Previous research had suggested that a familiar competitor placed against a novel referent under FM learning conditions could generate immediate lexical engagement, as the competitor activated a schema shared with the novel referent (Coutanche & Thompson-Schill, 2014, 2015; Coutanche & Koch, 2017; Coutanche, 2019). Demonstrations that schema support memory and learning are present elsewhere in the literature also (Havas et al., 2018; McClelland, 2013; Tse et al., 2007).

However, there were some conceptual problems with the FM literature, and many unanswered research questions (Section 5.1, p. 55). Moreover, the understanding of FM in the above memory literature was not reflected by antecedent developmental literature, from which the procedure was adapted (e.g., Carey & Bartlett, 1978). In the developmental literature, the FM task rarely makes reference to a feature shared between a competitor and the novel referent (e.g., Dysart et al., 2016; Riggs et al., 2015). Experiment 1 therefore intended to extend the FM findings by adapting the task to make it more like the developmental procedures. In doing so, it was also able to test the contention that the competitor alone was responsible for the accelerated emergence of lexical engagement (Coutanche & Thompson-Schill, 2014, Experiment 2) and, theoretically, schema activation. Instead of asking a question like "Are the antennae of the 'fostil' pointing up?" – a question which made explicit reference to a feature shared by the novel referent and familiar competitor – Experiment 1 used the question, "Where is the X?", which did not require participants to code the presence of absence of a shared feature, only a spatial location (left/right of the screen). Under these conditions, no evidence of lexical competition was observed. This is consistent with research carried out questioning the veracity of the FM effect (Cooper et al., 2019a, 2019b, 2019c; Gaskell & Lindsay, 2019; O'Connor et al., 2019; O'Connor & Riggs, 2019). However, training accuracy was above chance, as was later recognition accuracy, suggesting the results were not due to poor learning.

It is important to stress that at the time that this experiment ran, none of the above work doubting an FM effect had been published. Furthermore, it remains the case as of September 2021 that no published work has failed to replicate Coutanche and Thompson-Schill's (2014) work (though see Cooper et al., 2019a; Gaskell & Lindsay, 2019), and the effect has been replicated in a published paper (Coutanche & Koch, 2017; see also Zaiser et al., 2019b). Thus, it is valid to ask: are there other reasons that could have meant that Experiment 1 found no evidence for immediate lexical integration, leaving aside the possibility that the effect may not replicate at all?

[7] $d = \frac{M_1 - \frac{1}{3}}{SD_1}$; Cohen (1988). This formula (or its variant) will be used throughout this thesis for one or independent samples tests unless otherwise specified.

The first thing to emphasise is that participants did not appear to struggle with learning through FM generally, and it is only the claim that FM produces better or faster learning that needs to be interrogated. Consistent with work arguing both for and against FM's ability to promote immediate integration, recognition performance was above chance (e.g., Cooper et al., 2019a; Coutanche & Thompson-Schill, 2014). This suggested that participants did learn the novel words, but that these representations were not integrated sufficiently with their known competitors to delay their recognition in the semantic categorisation task.

Another possibility to discount is a methodological problem resulting from a lack of power. Given otherwise fixed parameters, power is a function of the number of participants, and also, the number of items those participants respond to. Whilst the number of items was reduced by 25%, the sample of participants was 50% larger than that used by (Coutanche & Thompson-Schill, 2014, $N = 25$, but $N = 37$ here). Furthermore, with respect to the items, even when a participant showed an RT difference, this difference did not correspond to a decrease in accuracy – suggesting that the RT difference was not the result of lexical competition (cf., Bowers et al., 2005).

Given the absence of obvious reasons why no lexical engagement may have been observed, two possibilities remain. The first is that the rapid lexicalisation of words trained by FM is a true effect, but sensitive to a narrow set of very particular conditions. The second is that the reported effect is not true. With the present experiment, it was impossible to distinguish between these, which needed to be addressed. Thus, Experiment 2 was a methodological replication of Coutanche and Thompson-Schill (2014), to provide evidence of replicability one way or another. Without such evidence, the interpretation of the results in Experiment 1 remained very difficult.

CHAPTER

SIX

EXPERIMENT 2
REPLICATING FAST MAPPING EFFECTS

6.1 Introduction and rationale

A limitation of Experiment 1 was that without knowing if the effects reported by Coutanche and Thompson-Schill (2014, see also Coutanche & Koch, 2017) were replicable, interpretation of Experiment 1 was very difficult. Coutanche and Thompson-Schill had three main findings:

1. that lexical engagement emerged 10 minutes after training for words trained by fast mapping (FM) only;

2. that this faster lexicalisation was at the cost of weaker declarative memory as measured by a three-alternative forced choice task (3-AFC), relative to a condition training words by explicit encoding (EE);

3. that lexical engagement as measured by semantic priming again emerged faster for FM rather than EE-trained words, but not until a second day of testing.

Furthermore, as Coutanche and Thompson-Schill had found in their second experiment that the question in the FM task was not enough on its own to bring about these FM effects, they concluded that it was an unimportant aspect of the task – and instead that the familiar competitor was activating a schema which allowed the immediate integration of novel information (cf., Tse et al., 2007; McClelland, 2013). Although they do not look for these FM effects, the developmental literature from which FM is borrowed likewise does not include such questions (e.g., "Are the antennae of the 'fostil' pointing up?", cf., Carey & Bartlett, 1978). Consequently, Experiment 1 changed the question from one requiring a semantic mapping (i.e., one must encode that a 'fostil' has antennae to answer a question about its antennae) to one requiring only a spatial mapping ("Where is the fostil?" requires only encoding that the 'fostil' is on the right or left to answer correctly). If under these conditions one had still found an FM effect, then one would be able to argue strongly that the familiar competitor was indeed supporting lexicalisation. This would have led to further experiments dissecting this effect.

Unfortunately, however, no such effect was found. Under such circumstances, one of two possibilities seemed likely:

1. that the 'FM effect' only operated under a very strict set of conditions, and disruption to these caused the effect to become undetectable in Experiment 1.

2. that Coutanche and Thompson-Schill produced a false positive, further replicated by Coutanche and Koch (2017), and that there is in fact no true 'FM effect' (perhaps with the exception of weaker declarative memory);

Experiment 2 was a methodological replication of Coutanche and Thompson-Schill (2014), run to distinguish between these possibilities.

6.1.1 The comparison to Coutanche and Thompson-Schill (2014)

To bring Experiment 2 into line with Coutanche and Thompson-Schill (2014), several changes were made to the design of Experiment 1. Firstly, as in Coutanche and Thompson-Schill's work, Experiment 2 took place over two days. This allowed for the tracking of the consolidation of the newly learnt words. Secondly, the EE condition was re-included, having been dropped in Experiment 1, to act as a base line to the FM condition. An important part of the findings in the FM literature is that encoding by FM comes with some sort of trade off (see Chapter 4, p. 41). Finally, the FM carrier question was that used by Coutanche and Thompson-Schill ("Are the antennae of the 'fostil' pointing up?"), and not as in Experiment 1. The EE condition used the same instruction as in their work also: "Remember the fostil".

As it had no obvious effect in Experiment 1, to be methodologically closer to Coutanche and Thompson-Schill, the familiar catch trials were removed from the training task. However, the number of objects learnt was held at 12. This was done as the 3-AFC performance in Experiment 1 was not particularly strong, and to allow a better comparison with between effects observed in Experiments 1 and 2. For the same reason, and to keep the experiment as short as possible, the semantic priming task (seen in Coutanche and Thompson-Schill's first experiment) was still not included. In all other ways the experiment was a full methodological replication, contrasting between EE with FM groups, over two days.

Lastly, a note on predicted results, as there are competing claims, even within the FM literature. Complete replication of Coutanche and Thompson-Schill (2014) would have meant the detection of an exposure × trial interaction, where FM exposure led to a positive former hermit − hermit difference, but no such difference for EE participants. In their paper, no day effect was observed as this pattern persisted overnight; however, Himmer et al. (2017) reported a consolidation effect for EE but not FM words, purportedly related to the faster integration of FM memories. However, Walker et al. (2019) reported no consolidation (or, indeed, competition) for words learnt with so few exposures. Another point of confusion was whether an effect would emerge on the lexical engagement task in either, or both, of the accuracy and response time (RT) data sets (compare Bowers et al., 2005; Coutanche & Thompson-Schill, 2014).

6.2 Methods

6.2.1 Participants

Technical and recruitment problems meant that the number of participants analysed in Experiment 2 was somewhat reduced. Out of around 90 participants tested, due to a combination of participants not returning and missing data (e.g., mis-recording, computer crashes), complete data was only easily extractable for 58 participants (eight male, M_{age} = 19.8 years, SD_{age} = 2.62 years). This was still gave a sample larger than that tested by Coutanche and Thompson-Schill (in their first experiment, $N = 50$). All participants had not participated in Experiment 1. All were free of any confounding disorders (e.g., sensory, learning or language difficulties), or had corrections to normal (e.g., by wearing eyeglasses).

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, or in exchange for course credits.

6.2.2 Materials and apparatus

Materials were similar to those used in Experiment 1, and used the same stimuli set (see Fig. A.1 and Table A.1, p. 213), although with altered training carrier phrases.

6.2.3 Design

As in Experiment 1, participants were randomly assigned to one of two lists of words (see Table A.1, p. 213). Additionally, participants were assigned to one of two exposure types: EE or FM. Of the 58 participants whose data were analysed, 29 were on each list, of whom 14 were assigned to EE exposure, and 15 to FM exposure. There were therefore 28 EE participants and 30 FM participants.

Day was manipulated within-subjects to test for consolidation of the newly-learnt words. Participants on the first day of training ('Day 1') completed training, then the lexical engagement task (semantic categorisation; Bowers et al., 2005), then the lexical configuration task (a 3-AFC). On the second day ('Day 2'), participants again completed the two lexical tasks in the same order as before, and then completed the familiarity questionnaire (as in Experiment 1 and Coutanche and Thompson-Schill), to check for pre-experimental familiarity with the novel referents.

As before, the lexical engagement task took responses to words in one of three trial types: fillers, former hermits and hermit words. Hermit words had no orthographic neighbours (Bowers et al., 2005), but transitioned to former hermits by participants learning substitution competitors (e.g., 'amazon' → 'alazon'). Whether a word was a hermit or not varied with list between groups of participants.

The experiment therefore had three independent variables: day (Day 1, Day 2) and trial (fillers, former hermits, hermit words) – both manipulated within subjects – and exposure (EE, FM) – manipulated between groups.

6.2.4 Procedure

Participants began on Day 1 with training, during which they saw each novel referent twice, with text printed under it. If participants had been assigned to the FM condition, the novel referent was in the presence of a same taxonomic class competitor and a question introducing the novel word (e.g., "Are the antennae of the 'fostil' pointing up?"); in EE, the novel referent was alone with an instruction to "Remember the X", where 'X' was the novel word, intended to be mapped to the novel referent. In the EE condition, no response was required; in the FM condition, participants selected the referent for the novel word by keypress. Accuracy and RT data were recorded, but only accuracy data were analysed. Regardless of when a participant responded, the trial was held on screen for 6s.

The rest of the tasks were identical to Experiment 1 (Section 5.2.4, p. 59). After training, participants watched a short video (∼10 minutes), to suppress rehearsal and introduce a retention delay, before completing the semantic categorisation task. Here, participants made 48 natural/man-made judgements to words printed on screen, divided across three trial types (filler, former hermits, hermits; see Table A.1, p. 213). Accuracy and RT data were analysed. Lastly, participants completed a 12 trial 3-AFC. Here, participants had to respond with one of three keys to indicate the presence of the correct referent for an on-screen word on either the left, right or centre of the screen. Each referent appeared as the foil for two other referents. Only accuracy data were analysed.

Participants came back at the same time the next day to complete the experiment. On Day 2, participants again completed the lexical tasks (semantic categorisation and 3-AFC), in exactly the same manner as on Day 1. Lastly, they scored each referent on its familiarity before the experiment began, on a 7 point Likert scale, as in Experiment 1 and Coutanche and Thompson-Schill (2014). This task was paper based, but the training, lexical engagement and lexical configuration tasks were scripted in DMDX (Forster & Forster, 2003).

Processing of the data and exclusions

Exclusions were processed as in previously published research (Bowers et al., 2005; Coutanche & Thompson-Schill, 2014), and as in Experiment 1.

Training. Compared to Experiment 1, a relatively large amount of training trials were removed. EE trials were not analysed as no response was required. Participants in the FM condition seemed to make very slow responses relative to Experiment 1: 20.3% of trials were excluded due to a response not being recorded within 6s. Moreover, participants frequently made an incorrect response: a further 24.5% of trials were incorrect.

This posed a problem, as the number of excluded training trials was very high, and far in excess of that in Experiment 1 (where the minimum accuracy was 87.5%). Unfortunately, Coutanche and Thompson-Schill (2014) did not report their training accuracy figures, so it is difficult to make a true like-for-like comparison, as the training tasks were different between Experiments 1 and 2. The problem did not appear to be confined to particular participants (see Section 6.3, p. 69). However,

a decision was taken not to exclude individual participants, as the same could not be done for the EE condition. This would have had the effect of biasing the FM condition, by selecting only the best responders. Moreover, in the case of time-outs, the participant's response may not have been incorrect (just slow), and due to the fixed length of each training trial, it was not the case that they received additional exposure from slower responding. In all cases, regardless of response accuracy, participants received 6s exposure per trial. Lastly, even for an incorrect response, training may have been sufficient to bring about the effect regardless (thus weakening the rationale for exclusion still further): the central measure of interest was the lexical engagement measure, and if FM functioned as described by Coutanche and Thompson-Schill (2014), its presence alone should have caused the automatic integration between novel referent (e.g., 'fostil') and familiar competitor (e.g., 'fossil').

Lexical engagement. Trials were excluded on the basis of incorrect responses (∼11% of all trials), RT < 300ms (no further removals), and RT > 1500ms (∼7% of all trials). Using the familiarity check task delivered at the end of testing on Day 2, former hermit trials where the novel referent was not unfamiliar (familiarity > 3; ∼3% of all trials) were also excluded. Subjects were then excluded on the basis of rating more than half of their learnt referents as familiar (none excluded), or due to having less than 70% of their trials remaining (total: 11 subjects). For the remaining participants, this meant they contributed an average of 40.6 trials (out of a maximum of 48; i.e., 84.6% of their trials).

Lexical configuration. Trials were excluded due to an incorrect (∼46% of all trials) or anticipatory (RT < 300ms: < 1% of all trials) response. As in Experiment 1, participants excluded from the lexical engagement analyses were allowed to contribute lexical configuration data.

6.3 Results

All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016).

6.3.1 Training

Training performance was similar across each of the training lists (M_1 = 70.6%, SD_1 = 12.0%; M_2 = 67.7%, SD_2 = 16.5%), and statistically identical (Welch's (1947) two sample t-test due to unbalanced groups: $t(16.3) = 0.448$, $p = 0.660$, $d = 0.197$). On this basis, all further comparisons collapsed across training lists.

Response accuracy was also found to be above chance ($t(20) = 0.628$[1], $p < 0.001$, $d = 1.37$). This confirmed that even with many trials removed, participants were not responding randomly, and therefore, the rest of the analyses could proceed.

[1]Note that the degrees of freedom reduction comes from the trial by trial exclusion of participants only – no participants were intentionally completely excluded.

Table 6.1 Summary of lexical engagement descriptive statistics in Experiment 2

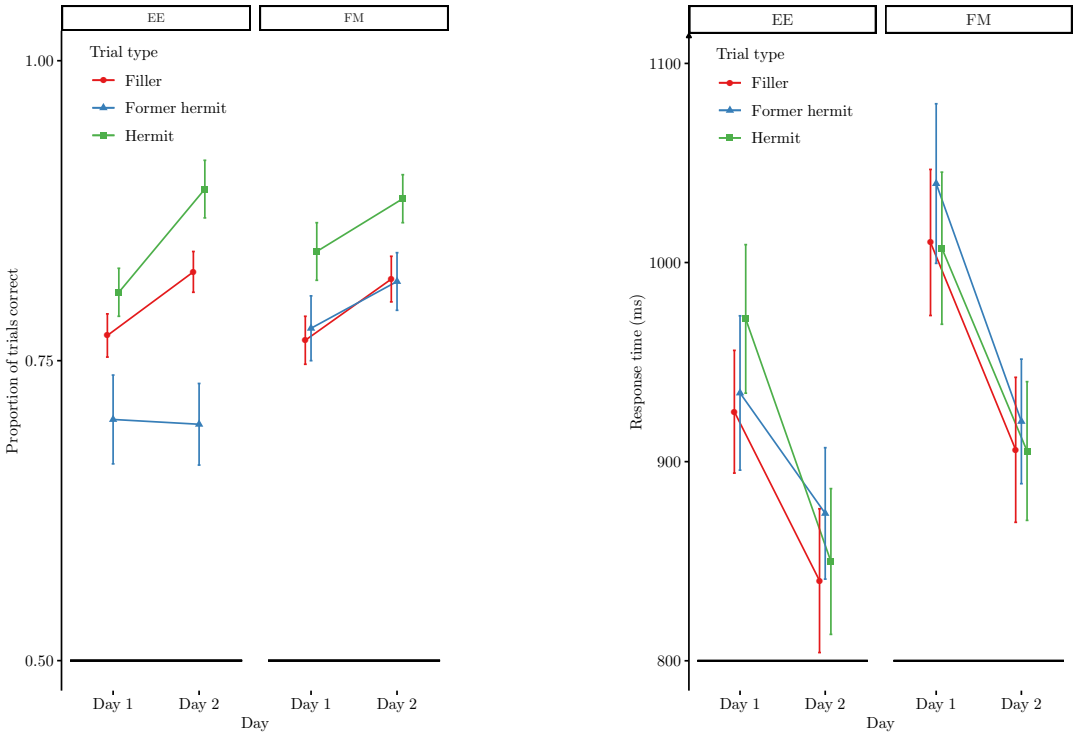
Accuracy data (% correct responses)					
Exposure	**Trial**	**Day**			
		Day 1		*Day 2*	
		M	*SD*	*M*	*SD*
EE	*Filler*	77.1	9.50	82.4	8.80
	Former hermit	70.1	19.8	69.7	17.8
	Hermit	80.7	10.8	89.3	12.5
FM	*Filler*	76.7	10.8	81.8	10.2
	Former hermit	77.7	14.9	81.6	13.3
	Hermit	84.1	13.0	88.5	10.8
RT data (ms)					
EE	*Filler*	925	163	840	191
	Former hermit	934	205	874	175
	Hermit	972	197	850	194
FM	*Filler*	1010	201	906	199
	Former hermit	1040	219	920	171
	Hermit	1007	209	905	191

6.3.2 Lexical engagement

Descriptive statistics summarising the accuracy and RT scores across the three independent variables of day (Day 1, Day 2), exposure (EE, FM) and trial (filler, former hermit, hermit) are displayed in Table 6.1 (p. 70), visualised in Fig. 6.1 (p. 71). Accuracy and RT data were separately subjected to mixed 2 × 2 × 3 ANOVAs, inputting the three IVs (Tables 6.2 and 6.4, pp. 71 and 73). Post-hoc *t*-tests were performed as appropriate (Tables 6.3 and 6.5, pp. 72 and 73).

Lexical engagement accuracy data

In the accuracy data, there were main effects of day and trial, and a significant day × exposure × trial interaction (Table 6.2, p. 71). Additionally, a main effect of exposure was very close to significance ($p = 0.051$), as was an exposure × trial interaction ($p = 0.080$).



(a) Accuracy data

(b) RT data

Figure 6.1: Lexical engagement condition means plot, on accuracy and RT data, for each day, exposure, and trial type. Errors bars show ±1 *SE*.

Table 6.2 Summary of lexical engagement accuracy ANOVA in Experiment 2

Effect	***F***	***p***	η_g^2
Day (1, 45)	22.9	$< 0.001^{***}$	0.044
Exposure (1, 45)	4.04	0.051, *NS*	0.021
Trial (1.34, 90)	13.8	$< 0.001^{***\dagger}$	0.124
Day × *Exposure* (2, 90)	0.397	0.532, *NS*	0.001
Day × *Trial* (1, 45)	1.51	0.226, *NS*	0.007
Exposure × *Trial* (1.34, 90)	2.95	0.080, $NS^{\dagger}$	0.029
Day × *Exposure* × *Trial* (2, 90)	3.48	0.035^{*}	0.016

Note. df given after the effect. Three asterisks (***) denotes significance below the 0.001 level; one asterisk (*) below the 0.05 level. † indicates Greenhouse-Geisser corrected-p value, due to non-sphericity (ε= 0.670; $W = 0.508$, $p < 0.001$)

Table 6.3 Summary of lexical engagement (hermit − former hermit) accuracy t-tests in Experiment 2

Exposure	Day	t	p	d
EE	*Day 1*	1.86	0.076, *NS*	0.579
	Day 2	4.41	$< 0.001^{*}$	1.37
FM	*Day 1*	1.98	0.059, *NS*	0.586
	Day 2	1.93	0.065, *NS*	0.506

Note. *df*: 21 for EE; 24 for FM. An asterisk (*) denotes significance below $\alpha = 0.013$, due to the Bonferroni correction.

In planned comparisons, lexical competition was then tested for by comparing former hermits against hermits, separately for each day and exposure type, giving four comparisons. A significant result in this block of four tests would be suggestive of lexical engagement, as the novel word disrupted responding to former hermit words more than hermits, making responses to them less accurate. Table 6.3 (p. 72) summarises the post-hoc paired-samples t-tests on the accuracy data. A significant difference was only observed on Day 2 for words learnt by EE.

Lexical engagement RT data

As with the accuracy data, the RT data were subjected to a 2 × 2× 3 mixed ANOVA, entering the same variables as before. This showed, again, main effects of day and trial, but no other effects or interactions. Exposure was however only marginally non-significant (results summarised in Table 6.4 (p. 73).

The lack of any interaction effects meant that Experiment 2 failed to replicate Coutanche and Thompson-Schill (2014). However, in order to investigate this further, the same battery of t-tests was performed on the RT data as was done for the accuracy data. These paired t-tests looked for evidence of competition, separately on Day 1 and Day 2 for EE and FM (summarised in Table 6.5, p. 73). The tests showed no evidence of competition for EE or FM exposed words, on either day, confirming the failure to replicate.

Lexical engagement supplementary analyses

As in Experiment 1, it may have been the case a drop in the number of items from 16 to 12 reduced the power in Experiment 2 to detect an effect. A supplementary analysis was carried out in Experiment 1 (p. 61), looking only at those participants who showed a positive former hermit − hermit RT difference; this would normally be indicative of a lexical competition effect. If this difference was truly indicative of a lexical competition effect, one would expect it to correlate to other indicators of competition, such as decreased accuracy (cf., Bowers et al., 2005). The size of the competition effect was therefore expected to correlate negatively with accuracy, if novel competitors were truly engaging the familiar words in the minds of a subset

Table 6.4 Summary of lexical engagement RT ANOVA in Experiment 2

Effect	F	p	η_g^2
Day (1, 45)	31.6	$< 0.001^{***}$	0.085
Exposure (1, 45)	3.04	0.088, *NS*	0.044
Trial (2, 90)	8.51	0.001^{***}	0.020
Day × *Exposure* (2, 90)	1.15	0.290, *NS*	0.003
Day × *Trial* (1, 45)	1.21	0.303, *NS*	0.002
Exposure × *Trial* (2, 90)	0.741	0.479, *NS*	0.002
Day × *Exposure* × *Trial* (2, 90)	0.567	0.569, *NS*	0.001

Note. df given after the effect. Three asterisks (***) denotes significance at or below the 0.001 level

Table 6.5 Summary of lexical engagement (former hermit − hermit)RT t-tests in Experiment 2

Exposure	**Day**	t	p	d
EE	*Day 1*	0.163	0.872, *NS*	0.034
	Day 2	1.03	0.314, *NS*	0.171
FM	*Day 1*	1.71	0.100, *NS*	0.228
	Day 2	0.359	0.722, *NS*	0.069

Note. df: 21 for EE; 24 for FM. An asterisk (*) denotes significance below $\alpha = 0.013$, due to the Bonferroni correction.

Table 6.6 Supplementary analysis correlating a participant's former hermit trial accuracy to their former hermit − hermit trial RT difference in Experiment 2, for participants with a positive difference only

Cell	Difference (ms)		$\text{Accuracy}_{\text{fh}}$ (%)		r	p
	M	*SD*	*M*	*SD*		
EE: day 1	81	77	75.0	22.2	−0.252	0.385, *NS*
EE: day 2	61	45	73.1	20.0	−0.559	0.074, *NS*
FM: day 1	79	62	84.3	9.50	0.170	0.500, *NS*
FM: day 2	63	50	85.0	10.5	−0.150	0.594, *NS*

Note. *N*s as follows. EE day 1: 14; EE day 2: 11; FM day 1: 18; FM day 2: 15. Pearson's r reported.

of the sample. The same calculation was performed here, separately for each day and exposure type. A summary of this analysis can be seen in Table 6.6 (p. 74).

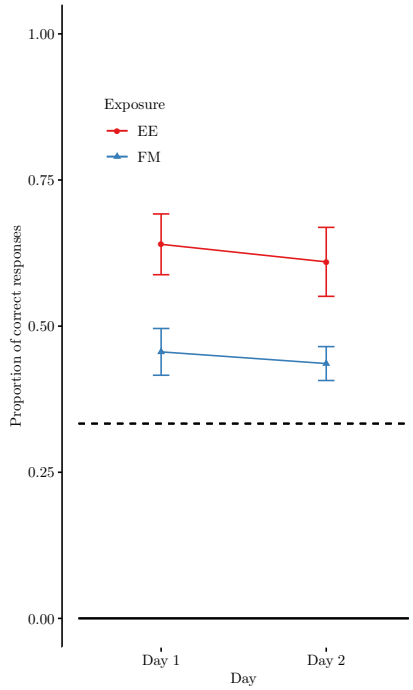
In the FM group there were no significant correlations on either day, whereas in the EE group there was a near-significant effect on Day 2 only. Thus, these correlations confirm the general pattern shown in the accuracy and RT data: the strongest evidence for any lexical competition effects were seen only on Day 2 and in the EE group, contrary to the findings of Coutanche and Thompson-Schill (2014).

6.3.3 Lexical configuration

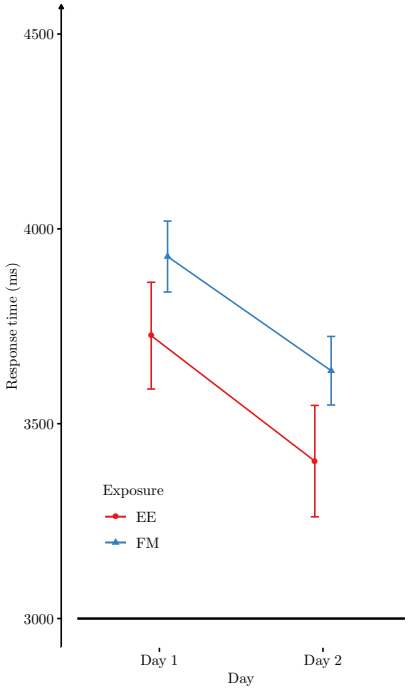
A summary of the 3-AFC task data can be seen in Fig. 6.2 (p. 75). On average, participants memory was best for items learnt by EE on the first day of testing ($M = 64\%$ recalled, $SD = 27.4\%$), and worst on the second day of testing for items learnt by FM ($M = 43.6\%$, $SD = 16.0\%$). As shown in Fig. 6.2a, both groups experienced forgetting over the two days of the experiment, with numerically worse performance on the second day. Also, recognition was consistently numerically higher for words learnt by EE. This is consistent with much research on both sides of the FM debate (e.g., Cooper et al., 2019c). Participants were also consistently slower in the FM condition, but both groups did show RT improvements, with responses being numerically faster on the second day ($M_{difference\ EE} = 322$ms; $M_{difference\ FM} = 293$ms).

Lexical configuration accuracy data

The accuracy ANOVA showed a main effect of exposure, but no other main effect, and no interaction (Table 6.7, p. 76). Accordingly, data were collapsed across Day 1 and Day 2, and post-hoc t-tests were then performed on this data set. Three comparisons were made with the accuracy data: EE against chance level performance, FM against chance level performance, and EE against FM (Welch's t-test, due to unequally sized samples). These tests showed that whilst both EE and FM led to above chance recognition accuracy, EE resulted in superior recognition (Table 6.8,



(a) Accuracy data. Chance performance indicated by dashed line (= 0.333)



(b) RT data

Figure 6.2: Data from the 3-AFC task. Errors bars show ±1 *SE*.

p. 76). This result is consistent with findings elsewhere in much of the literature (for review, see Cooper et al., 2019c).

Lexical configuration RT data

The RT ANOVA showed a main effect of day, but no effect of exposure, nor an interaction (Table 6.7, p. 76). The data were therefore collapsed across exposure types. Responses on Day 1 were found to be significantly slower than responses on Day 2 (paired-samples t-test; $t(57) = 3.42$, $p = 0.001$, $d = 0.489$).

6.4 Discussion

Experiment 2 attempted to replicate the results of Coutanche and Thompson-Schill (2014), in order to interpret the lack of a lexical engagement effect in Experiment 1. These researchers had found that lexical engagement emerged faster under FM learning conditions than it did under EE conditions. However, they also found that this was at the cost of weaker declarative memory, as words learnt by FM were less well recognising on a 3-AFC task. Unfortunately, across several measures, there was no indication of lexical engagement in Experiment 2. However, there was weak evidence of consolidation in the 3-AFC task, and FM did produce weaker memory traces than EE. The findings are summarised below.

Table 6.7 Summary of 3-AFC accuracy and RT ANOVAs for the lexical configuration (3-AFC) task in Experiment 2

Accuracy data			
Effect	$\boldsymbol{F}$	$\boldsymbol{p}$	$\boldsymbol{\eta_g^2}$
Day	0.540	0.466, *NS*	0.003
Exposure	10.5	0.002^{**}	0.120
Day × *Exposure*	0.024	0.878, *NS*	< 0.001
RT data			
Day	11.5	0.001^{***}	0.059
Exposure	2.53	0.117, *NS*	0.030
Day × *Exposure*	0.025	0.875, *NS*	< 0.001

Note. $df = (1, 56)$. Three asterisks (***) denotes significance at the 0.001 level; two asterisks (**) below the 0.01 level.

Table 6.8 Summary of 3-AFC accuracy post-hoc *t*-tests in Experiment 2

Accuracy data			
Comparison	$\boldsymbol{t}$	$\boldsymbol{p}$	$\boldsymbol{d}$
EE vs. chance (27)	6.30	$< 0.001^{*}$	1.19
FM vs. chance (29)	3.59	0.001^{*}	0.655
EE vs. FM (47.9)	3.21	0.002^{*}	0.849

Note. *df* given after the comparison. An asterisk (*) denotes significance below $\alpha = 0.017$, due to the Bonferroni correction.

6.4.1 Summary of the lexical engagement findings

Both accuracy and RT measures showed main effects of day (i.e., between Days 1 and 2), and trial (i.e., between fillers, former hermit words – for which a competitor had been trained, and hermit words – for which a competitor had not been trained). Likewise, both measures showed a marginally non-significant effect of exposure (EE or FM). However, post-hoc tests on the RT data showed that the main effect of trial was not due to lexical competition, as comparing former hermit words to hermit words showed no difference in either of the EE or FM exposure groups, on either day.

The accuracy data did reveal a significant interaction between day, exposure and trials. However, this appeared to be driven by a lack of any competition effects in the FM condition on either day. In contrast, EE did show a difference between former hermit and hermit words on Day 2, but not on Day 1, although it is possible that this does not denote lexical competition. Examining the descriptive data, it is clear that the former hermit accuracy remained roughly flat (-0.4% in performance), whilst the hermit accuracy increased. A stronger demonstration of competition would have been changes in responses to the former hermit words, and an accompanying effect in the RT data. Given this lack, it seems possible at least that the difference was not caused by lexical competition.

It was also notable that the other post-hoc t-tests in the accuracy data produced only marginally non-significant results (all ps 0.059 – 0.076; all ds 0.506 – 0.586; $\alpha = 0.013$). In the context of this data alone, this might have suggested a true effect, but some problem in the experiment, for example, with power. However, it must be emphasised that no such marginal effect appears in the RT data, and the failure to replicate sits with other failures referenced in the literature (Cooper et al., 2019a; Gaskell & Lindsay, 2019), and Experiment 1. This can therefore be declared unlikely. Certainly, there is no *clear* demonstration of a lexical competition effect as seen in Coutanche and Thompson-Schill's work, and so the replication must be considered a failure.

Emphasising this further is the supplementary analysis that was performed, tying together the RT and accuracy data. One methodological change from Coutanche and Thompson-Schill was a drop in the number of word learning trials from 16 to 12. This change would have reduced power, making the detection of an effect more difficult. To sidestep this, participants apparently displaying the effect were tested alone. Subsetting the participants so that only those with a positive former hermit − hermit word difference were tested, this difference was then correlated to their response accuracy for the former hermit words. If the RT difference was driven by lexical competition, that increase in the difficulty of processing novel words should have made responses to the former hermits more inaccurate also, resulting in a negative correlation. However, all correlations were non-significant, except for the EE group on Day 2. This trend towards a competition effect in EE on Day 2 is more consistent with a complementary systems account, not the 'FM-effect' (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995).

6.4.2 Summary of the lexical configuration findings

Just as with the lexical engagement data, the 3-AFC task measuring lexical configuration gave contradictory indications, as there was disagreement between accuracy and RT measures. Whilst the accuracy data suggested that there was no main effect of day, it did suggest a difference between EE and FM. However, the RT data suggested the opposite: no difference between EE and FM, but a difference across days. Neither measure suggested an interaction.

Collapsing over day, the post-hoc *t*-tests on the accuracy data suggested that learning was possible in both the EE and FM conditions, as accuracy in both groups was above chance. This fits with accounts of word learning that argue the cognitive system is fairly flexible with respect to how training of the words takes place (e.g., Davis & Gaskell, 2009; Dumay et al., 2004; Henderson et al., 2013; Kapnoula et al., 2015). Participants were also found to be more accurate for words learnt by EE than for words learnt by FM, again, fitting with accounts from proponents and opponents of an 'FM-effect' (e.g., Coutanche, 2019; Warren & Duff, 2019).

The RT data showed that responses on Day 1 were significantly slower than on Day 2. A simple explanation for this finding is task familiarity: participants on Day 2 had become more practised at responding. The alternative explanation is that this is evidence of consolidation. However, given the lack of evidence for consolidation in accuracy data (i.e., no effect of day), it seems unlikely that this is a consolidation effect.

6.4.3 Conclusions, and future work

The conclusion to emphasise from Experiment 2 is that, according to these data, the 'FM effect' described by Coutanche and Thompson-Schill (2014) does not replicate. There is no evidence of immediate lexical competition in this experiment. On the surface, this would support a complementary learning systems account (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995).

With respect to the FM literature, Experiments 1 and 2 fit more closely with opponents of the purported FM effect. These researchers appear to have a plurality in the field (see Table 4.1, p. 53; Cooper et al., 2019d). Proponents of the effect have made various claims – for example, that FM might help with patients' memory (e.g., Atir-Sharon et al., 2015; Merhav et al., 2014, 2015; Sharon et al., 2011), or with faster lexicalisation in healthy populations (Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017). However, there is more evidence against the effects, particularly with respect to patients (Cooper et al., 2019b; Greve et al., 2014; Korenic et al., 2016; Sakhon et al., 2018; Warren & Duff, 2014; Warren et al., 2016). Although the fundamental idea – that schema support learning – may be sound (Havas et al., 2018; McClelland, 2013; McClelland et al., 2020; Tse et al., 2007), the specific argument that such schema are activated in FM and promote lexicalisation cannot be substantiated (cf., Cooper et al., 2019a; Gaskell & Lindsay, 2019). This aligns with the views of developmental researchers, who have been critical of FM's application and conceptualisation (Horst et al., 2010; O'Connor & Riggs, 2019; O'Connor et al., 2019; Vlach & Sandhofer, 2012).

However, the data in Experiments 1 and 2 are also a poor fit for a complementary learning systems account. Davis and Gaskell (2009), building on the conclusions of other work (e.g., Dumay & Gaskell, 2007; see also Tamminen et al., 2010), argued that a single night of sleep was sufficient to bring about behavioural change. The evidence for that behavioural change in this experiment was weak, and seems task and measurement dependent (see also McMurray et al., 2017; Palma & Titone, 2020). It may be that no systematic behavioural change was observed in this experiment due to the small number of exposures (Walker et al., 2019). However, this is a somewhat unsatisfactory explanation, as it implies that in some instance representations are *not* consolidated, with no clear explanation as to why. The idea of a representation's strength does not appear to take one far – after all, the representations were still strong enough to support recognition performance at above chance levels.

Instead, it may simply be that semantic categorisation, as used in Experiments 1 and 2, is a poor or insensitive measure of lexical engagement. Outside of whether specifically FM learning conditions can support early lexical engagement, alternative methodologies might be able to provide evidence for it, contrary to the predictions of complementary learning systems accounts (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995).

In the preceding chapters of this thesis, other work in the literature finding an immediate lexical competition effect has been referred to, but skipped over (for reviews, see McMurray et al., 2017; Palma & Titone, 2020). Part III of this thesis considers this literature and investigates the evidence for same-day lexical engagement further.

Part III

Lexical engagement in computer mouse tracking: optimising design factors in computer mouse tracking to detect immediate lexical engagement

CHAPTER

SEVEN

LEXICAL ENGAGEMENT BEFORE SLEEP

The first two parts of this thesis (Parts I and II) presented models and literature relating to speech perception, memory and word learning. The first experimental work, presented in Chapters 5 and 6, further explored experiments (e.g., Coutanche & Thompson-Schill, 2014) that had returned data incompatible with a complementary learning systems account of word learning (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). However, when extension (Experiment 1) and replication (Experiment 2) were attempted, these data were found not to replicate or extend. Whilst fast mapping (FM) did not lead to such immediate lexical engagement effects, may they be found elsewhere in the literature?

The short answer is 'yes'. There are several papers with paradigms other than FM showing evidence of pre-sleep lexical competition (e.g., Bartolotti & Marian, 2012; Fernandes et al., 2009; Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Lindsay & Gaskell, 2013; Szmalec, Page & Duyck, 2012; Weighall et al., 2017). The data from Experiments 1 and 2 are therefore only a repudiation of the fast mapping claims, and do not speak to 'rapid' (i.e., pre-sleep), or immediate, lexical engagement effects observed elsewhere, with other paradigms or measures. It should also be highlighted that although the focus of this thesis is lexical engagement *by means of lexical competition*, the bulk of the literature showing evidence of such rapid/immediate lexical engagement effects used paradigms other than lexical competition. However, the logic is the same: do newly acquired representations show evidence of possessing characteristics that could be said to be lexical?

The papers showing rapid/immediate lexical effects divide into two, and these groupings should be clearly distinguished. To be inconsistent with the complementary learning systems model (CLSM), an effect need not be *immediate*, merely *pre-sleep*. This is because sleep is thought to promote or cause consolidation (cf., Dumay & Gaskell, 2007), and consolidation is responsible for the lexicalisation of non-lexical traces of newly learnt words (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995). Whilst the focus here will be on immediate effects, some papers have detected lexical engagement pre-sleep, but not immediately (e.g., Lindsay & Gaskell, 2013).

This chapter will take the following shape. The first section will review pre-sleep

lexical engagement effects using paradigms beyond lexical competition, to demonstrate just how widespread these effects are. This will set a solid experimental base that will be useful for contextualising and review later in the general discussion (Chapter 15, p. 195). The second section of this chapter will set out the evidence for immediate and pre-sleep lexical *competition* effects specifically. This literature will then feed through to experimental work later in the thesis (Experiments 5–7; Chapters 12 to 14, pp. 155, 177 and 191, respectively).

7.1 Lexical engagement outside lexical competition

A very recent review of word learning emphasises that the model advocated by the authors such as Davis and Gaskell (2009) and Lindsay and Gaskell (2010) is no longer viable, as there is a "rich variety of time courses for novel word lexicalisation" (Palma & Titone, 2020, p. 1). Further work emphasises that engagement may be fast-emerging across many different aspects of lexical knowledge (e.g., phoneme-phoneme, lexeme-phoneme, lexeme-semantics; McMurray et al., 2017). A summary of this literature showing pre-sleep lexical engagement (excluding FM papers) is summarised in Table 7.1 (p. 85). A review of this body of work emphasises that effects may emerge very quickly indeed, as most papers show engagement in a testing session immediately after training. It is also present across a variety of measures, both behavioural (offline, e.g., Lindsay & Gaskell, 2013; and online, e.g., Weighall et al., 2017) and neuroscientific (e.g., Bakker, Takashima, van Hell, Janzen & McQueen, 2015).

Although not a word learning paper, Betts, Gilbert, Cai, Okedara and Rodd (2017) showed that for words with ambiguous meanings (e.g., 'bark' – relating to either a dog, or a tree), the meaning which participants prefer may be skewed by recent experience. This does not speak to how novel words may be handled by cognitive systems; however, this is evidence that recent experience may alter how stored, fully 'lexicalised' representations may be influenced by information processed presently. Likewise, it is possible that novel words may be similarly inter-connected with, and alter the processing of, known words in the period shortly following a learning episode.

Eye tracking evidence from Kapnoula and Samuel (2019) builds on this. Studying so-called indexical effects present in the speech signal (namely, the identity of a speaker), the authors trained novel words in one of three voices. The words were in L2, and referred to known concepts (e.g., 'bifa' meant 'kite'). The researchers manipulated the usefulness of the voice information: either, it was systematically paired with a particular referent (e.g., voice 1 uttering the word 'bifa' always mapped to the green kite; voice 2 saying the same word always mapped to the blue kite) or to no particular kite (both the blue and green kites appeared with each speaker). Participants were then faster to fixate on the target item for words which had been trained with a particular voice, showing that indexical information could become linked to semantic information, and that this link could emerge rapidly. This effect replicated in a second experiment, and a third experiment showed that the effect was not modulated by sleep. This speaks against a complementary learning systems account whereby episodic details are abstracted away as cortical systems generalise

Table 7.1 Summary of pre-sleep lexical engagement literature, excluding FM papers

Citation	**Semantic training /effect(s)?**	**Immediate effect(s)?**	**LCE?**
Bakker et al. (2015)	✓	✓	✗
Bartolotti and Marian (2012)	✓	✓	✓
Fernandes et al. (2009)	✗	✓	✓
Geukes et al. (2015)	✓	✓	✗
Kapnoula et al. (2015)	✗	✓	✓
Kapnoula and McMurray (2016a)	✗	✓	✓
Kapnoula and Samuel (2019)	✓	✓	✗
Laine et al. (2013)	✗	✓	✗
Leach and Samuel (2007)	✓	✓	✗
Lindsay et al. (2012)	✗	✓	✗
Lindsay and Gaskell (2013)	✗	✗	✓
Snoeren et al. (2009)	✗	✓	✗
Szmalec et al. (2012)	✗	✗	✓
Tham et al. (2015)	✓	✓	✗
Weighall et al. (2017)	✓	✓	✓

Note. LCE = Lexical competition effect.

across experiences. These data are important in demonstrating that: (1) the complementary learning systems distinction between 'episodic' and 'lexical' may not be meaningful – as lexical representations stored details of recent experience, such as speaker identity – and (2) emphasising that lexical links may emerge rapidly, and independently of sleep. This fits well with accounts doubting the existence of an abstracted and distinct mental lexicon (e.g., Dilkina, McClelland & Plaut, 2010), or at least, of episodic representations within the lexicon (e.g., Goldinger, 1998). Effects showing that participants code and use the identity of speakers in lexical processing have been demonstrated elsewhere in the literature also (Goldinger, 1996; Cai et al., 2017).

Other authors have also shown rapid lexical engagement effects in words trained with semantic meaning. Geukes et al. (2015) took the novel approach of applying the Stroop (1935) task to word learning. In this task, participants either see colour words written in a congruent colour (e.g., the word 'red' in red ink), or else in a different colour (the incongruent condition, e.g., the word 'red' in blue ink). When told to name the colour of the ink (e.g., by clicking on a box with that colour), and to ignore the word itself, a common finding is that participants give faster responses on congruent trials than on incongruent trials. This effect is apparently due to the automatic reading of a word, and the consequential activation of its semantic meaning, which then interferes with giving the correct colour response. Having learnt the novel words naming colours, and providing that novel word trials were intermixed with familiar word trials, Geukes et al. found that participants demonstrated a Stroop effect for novel names of colours immediately after learning. Note that this is not merely novel words engaging novel words (as in Magnuson et al., 2003) as the colour response options were not labelled with a novel form, and theoretically, participants could perform the task without reading the novel words at all. Therefore, the fact that a Stroop effect was present must have meant that automatic reading of the novel words occurred, that this activated semantics, which then interfered with making a response. This implies some linkage between familiar concept (e.g., the colour blue) and the novel word. Once again, this is incompatible with a CLSM which argues that such automatically activated links can only occur through slow consolidation.

However, when familiar word trials were not present, no such effect was found. Nevertheless, this is further evidence that under certain conditions, novel words may demonstrate rapid lexical engagement. Similarly, extenuating circumstances were present in Kapnoula and Samuel (2019)'s work: instead of the more common 10 or so exposures during novel word training (e.g., Walker et al., 2019), the researchers provided participants with 63 exposures per word, and acknowledged that whilst their data may not be representative of what humans do ordinarily (as words are rarely used only by one speaker), their work reflects a property of what the system *can* do. Geukes et al.'s work fits into this pattern also: with appropriate testing and or training procedures, novel word engagement may be observed very quickly, even if it is not detectable under all circumstances. For example, Leach and Samuel (2007) found that when words were trained with meaning, they were able to bias perceptions of phoneme categories. However, other authors have also shown that phoneme categorisation can be immediately biased, without training semantics (Lindsay et

al., 2012; Snoeren et al., 2009). Also at a sub-lexical level, Laine et al. (2013) showed that participants could generalise from the phonotactics of a learnt item set to correctly distinguish new items as belonging to the same artificial 'language' (see also Oh et al., 2020).

In summary, many papers have demonstrated effects, at a variety of levels (lexical, sub-lexical, morphological, etc., McMurray et al., 2017; Palma & Titone, 2020). Whilst the next section will consider effects at the lexical level, the evidence is so widely distributed that a CLSM view framing sleep as a *pre-requisite* for lexical engagement (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010) is no longer tenable (McMurray et al., 2017; Palma & Titone, 2020).

7.2 Pre-sleep lexical competition effects

Pre-sleep lexical competition effects were first shown with novel words by Magnuson et al. (2003). Although this was not taken as evidence of lexicalisation, and no test of novel-known word competition was made, those data demonstrated that novel words could acquire apparently 'word-like' properties very quickly. For example, the authors showed that even when a competitor was not present, it could alter the processing of a target. Similarly, neighbourhood density effects were observed (Luce & Pisoni, 1998). Novel words could also engage each other freely, as is commonly seen in the word learning literature – pairs such as 'dibu' and 'pibu' competed as cohort competitors.

Accepting the line of reasoning advanced by Gaskell and Dumay (2003), that the strongest demonstration of a novel word's lexical nature was it altering of responses to a known word, Fernandes, Kolinsky and Ventura (2009) used a statistical learning paradigm in an artificial language to demonstrate that streams of nonsense syllables could be lexicalised. Embedded within the streams were sequential syllables which could form 'words', as they occurred with high transitional probability (i.e., it was likely that syllable one would be followed by syllable two, then by syllable three, forming a 'word' of those three syllables). These high transitional probability syllable sets (from here on, novel 'words') were competitors for known (Portuguese) words. For example, if embedded and repeated in the stream of syllables were the tokens /fi/, /vɛ/, /ku/ (forming the word 'fiveku'), lexical decision latencies to a familiar competitor 'fivela' (/fiˈvɛlɐ/, 'buckle') were increased. This occurred immediately after hearing the syllable stream which played the novel words 189 times over 21 minutes, with 10 words to learn, intermixed with 1260 other syllables. The other syllables all had low transitional probabilities: whereas /vɛ/ always followed /fi/ (probability of 1), other syllables only had a one third probability of preceding/following another syllable. This replicated in a second experiment, but two other experiments showed that when the competing part of the novel word (i.e., the overlapping portion – in the example above, /fiˈvɛ/) was embedded inside another syllable set (e.g., /mu/, /fi/, /vɛ/) the effect was not present. This suggests that the syllables were not being processed as isolated as individual units, but like true words, and were lexicalised as a set (see also Dumay & Gaskell, 2012). This is consistent with the predictions of the distributed cohort model – the initial syllable /mu/ would not form a cohort with familiar words beginning /fi/, such as 'fivela', so no com-

petition would be observed between these forms (Gaskell & Marslen-Wilson, 1997). This was strong evidence both for immediate engagement, and also for a 'word-like' nature of the effect, as the novel words behaved similarly to familiar words. Although only showing an effect after 12 hours, and not immediately, Szmalec et al. (2012) also showed a pre-sleep effect for words trained in the same way as Fernandes and colleagues, but on a pause detection measure.

One possible reason for authors such as Fernandes et al. (2009) and Szmalec et al. (2012) showing competition effects, when others such as Dumay and Gaskell (2007) did not, is that the segmented syllables more strongly activated their familiar competitors. This would fit with recent updates to complementary learning systems theory (McClelland, 2013; Kumaran et al., 2016; McClelland et al., 2020). The update posits that information may be more rapidly incorporated where it is consistent with knowledge already stored, building on work on schema in animals (Tse et al., 2007; for example in human word learning, see Havas et al., 2018). Strong activation of the known word may allow integration of the new information – in a way similar to that articulated by authors in the fast mapping literature (e.g., Coutanche & Thompson-Schill, 2014; Merhav et al., 2014). Following this line of argument, Lindsay and Gaskell (2013) interleaved sessions of training and testing throughout the course of two days. In doing so, the authors were able to demonstrate a lexical competition effect on a lexical decision task within two and a half hours of initial exposure. Moreover, in later experiments, this effect was found not to be supported merely by spaced exposure to novel words, but further exposure to their known competitors was also required. This suggests that the co-activation or interleaving may support the activation of the novel word, allowing competition to emerge.

Further evidence for this is seen in the eye tracking literature. Weighall et al. (2017) used four training tasks. Firstly, participants would state aloud the novel word. Secondly and thirdly, they would segment the novel words, pronouncing first the initial, and then the final syllable. Lastly, they performed a two-alternative forced choice task with feedback. The segmentation may have resulted in learning which was functionally similar to that seen in the work of Fernandes et al. (2009) and Szmalec et al. (2012); isolated syllables might activate a cohort which included the familiar competitor. This co-activation of novel and familiar words may then have facilitated the rapid emergence of engagement.

Weighall et al. also discuss the possibility that their eye tracking task may be more sensitive to the activation of specific items (i.e., that links are formed between a novel word 'biscal' and 'biscuit', not between 'biscal' and all other cohort competitors, necessarily). This is in contrast to measures such as pause detection, which give a more general indication of the overall level of lexical activity (Gaskell & Dumay, 2003; Kapnoula & McMurray, 2016a; Mattys & Clark, 2002; Weighall et al., 2017). As such, a competition effect found by pause detection may represent a difficulty in processing an entire *cohort* of items, rather than specific interference between two items. Furthermore, as an offline measure, it does not allow the unfolding of competition over time to be noted – potentially masking an effect present only briefly. Thus, as the activation of a novel word can be more precisely measured with techniques such as eye tracking, competition is detectable sooner.

Weighall et al. (2017) found that when participants were trained as above on a novel word, fixations in a visual word paradigm (VWP; Tanenhaus et al., 1995) task were increased when the novel word was a cohort competitor for a target (i.e., when the novel competitor shared an onset syllable with the target), relative to a condition where the novel word was not a cohort competitor. This suggested that participants were automatically co-activating the competing novel form when they received input for the target. This also suggested that participants were not just looking to the novel object as they were familiar with it, having recently learnt the novel words. This effect occurred immediately after training, and persisted across a second day of testing (though did not strengthen or weaken). Additionally, it was present in both children and adults. Trials with two known cohort competitors (e.g., 'towel'–'tower') showed that the competition experienced by the novel words was similar, although Weighall et al. argued, qualitatively different, to that exhibited on trials with novel words. The work is robust evidence of immediate lexical engagement, but is also noteworthy as one of only two papers also training semantic referents and reporting immediate competition (the other being Bartolotti & Marian, 2012).

Kapnoula et al. (2015) also used VWP to study word learning and the emergence of lexical engagement by competition, although unlike Weighall et al.'s work, they did not look for their effect across time, or in children, or train a semantic referent. Instead, they trained a set of novel words, either 'explicitly' (using a task similar to Weighall et al.) or 'implicitly', through phoneme monitoring, in two separate experiments. Both experiments showed evidence of immediate competition, but the competition was measured in a way not seen elsewhere in the word learning literature (adapted from Dahan, Magnuson, Tanenhaus & Hogan, 2001). Participants learnt monosyllabic words which were cohort competitors for monosyllabic English words, e.g., 'jod' (/ʤɐd/), a competitor for both 'job' and 'jog'. The novel items all had two familiar competitors, creating item triplets of a single novel word and two known words.

Since phonological gradations are continuous, a function of the language system is to carve the continuous speech stream into distinct phonemes, where no physical distinction exists. For example, when pronouncing the vowel in these words, as the speaker transitions to the final stop consonant, there is a perceptible point of cross over; the word 'job' is pronounced such that from the vowel itself the word can be ascertained (the 'jo–' from 'job' to be represented here as /ʤɐᵇ/). This is a phenomenon called co-articulation.

These consonant-vowel tokens, by which the word itself could be recognised, were then spliced with the final consonant from the other familiar word in the triplet, creating the artificial form 'joᵇg' (/ʤɐᵇg/), with a co-articulatory mismatch. The effect of this was that the word 'job' was more strongly evoked when participants heard 'jog': a condition where the co-articulatory cues matched was used as a comparison condition ('joᵍg'; /ʤɐᵍg/). This resulted in lower fixations towards the target on these mismatched trials than on matching trials (see also Dahan et al., 2001). In a similar way, novel word lexical engagement could be tested: it was found that immediately after training novel words engaged in competition in exactly the same way.

To rule out the possibility that participants were simply unable to recognise

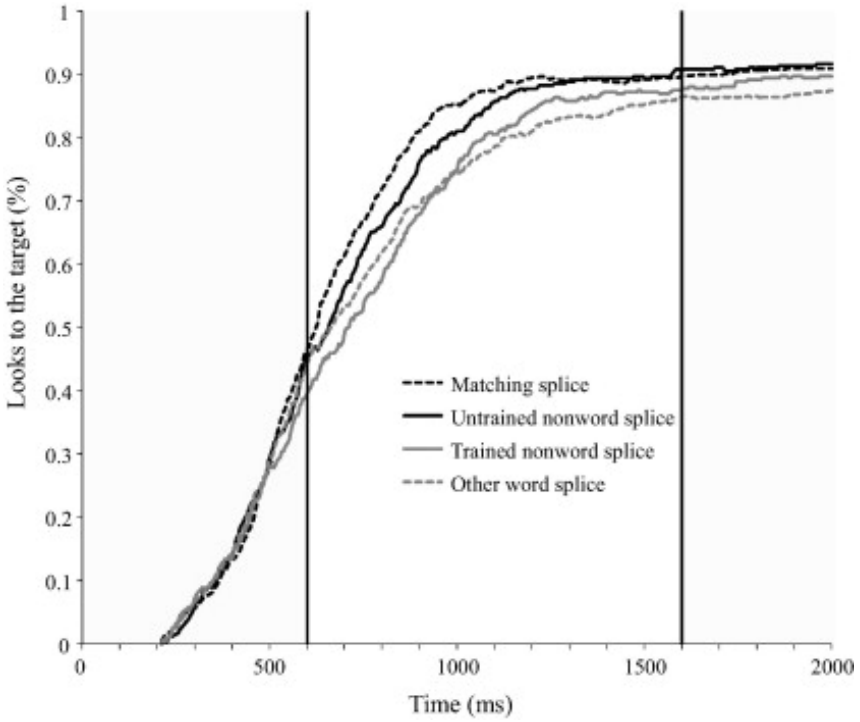


Figure 7.1: Data from Kapnoula et al. (2015, Experiment 1). Note: matching splice equivalent to 'jo^{g}g', evoking only 'jog'. Untrained non-word splice equivalent to 'ne^{p}t', having *not* learnt 'nep', thus giving a poor exemplar of 'net'. Trained splice equivalent to 'jo^{d}g', having learnt 'jod', and thus evoking that novel word. Other word splice equivalent to 'jo^{b}g', evoking 'job'. Trained/untrained difference significant; trained/other word difference non-significant, suggesting functionally equivalent familiar-familiar and novel-familiar word lexical engagement

the target word with the spliced stimuli, for half the triplets, the novel word was not trained. For example, whilst 'ne^{t}ck' still evoked 'net', having not learnt the novel word 'nep', 'ne^{p}t' was processed only as a poor exemplar was 'net' (with 'nep' evoked, but not recognised as a learnt item). Consequently, on these trials, fixations to the target were higher than when a learnt novel word was evoked. Having learnt 'jod', 'jo^{d}b' would evoke 'jod', and thus, prompt lexical engagement, shown by significantly lower fixations to 'job', the trial target. A graph summarising this finding can be seen in Fig. 7.1 (p. 90). Interestingly, the change in training from explicit to implicit did not result in a loss of the effects: this is presumably because the co-articulatory mismatch and VWP allowed for sufficient activation of the novel and familiar words at test for engagement to be measured, despite the fact that previous work had suggested that activation of the familiar word during training previously was important (e.g., Fernandes et al., 2009; Lindsay & Gaskell, 2013; Szmalec et al., 2012).

These findings were further built upon and extended by Kapnoula and McMurray (2016a). Using the same paradigm, and articulatory mismatch again, in this work, the effect was found to be robust to changes in speaker, suggesting that the novel words were indeed lexical representations, and participants were able to sufficiently

well generalise across different instances of words at test and training. This is, again, inconsistent with an account that portrays sleep-related consolidation as the process by which generalisation occurs (e.g., Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010).

In summary, much work in the literature shows evidence of pre-sleep or immediate lexical engagement, and it seems more likely that this is a property that emerges with appropriate testing and training regimes, not by exclusively by sleep or slow consolidation. However, sleep may still play in a role in strengthening and stabilising representations, and lexical configuration measures do typically show strong benefits of sleep (e.g., Dumay & Gaskell, 2007; McMurray et al., 2017; Palma & Titone, 2020).

7.3 Introducing mouse tracking

A final paper, using mouse tracking, is also noteworthy. Mouse tracking is a technique that functions similarly to eye tracking – indeed, participants often perform exactly the same task. In mouse tracking and in VWP, participants must click on targets in response to an instruction (e.g., 'Click on the X'). However, unlike eye tracking, which is technically difficult, and produces data that is essential binary – a participant is either fixating on an object, or not; a saccade is launched, or not – mouse tracking allows the smooth and gradated response of every single participant, on every trial, to be analysed, as the position of the mouse as it travels across the screen is logged (Spivey et al., 2005). This contrast with so-called 'ballistic' eye tracking has been considered an advantage of the technique, and may make it particularly suitable for studying novel word learning, as it allows specifically for the imaging of the processing of the unfolding speech signal and resultant lexical engagement (Bartolotti & Marian, 2012; Spivey et al., 2005).

In the first paper pairing mouse tracking and novel word learning, Bartolotti and Marian (2012) reported an effect consistent with the argument seen above. As the eye tracking research above, it was found that immediate novel word lexical engagement occurred under conditions where the activation of the novel word was sufficiently strong. Although no segmentation took place, participants were trained intensively to criterion (90% production accuracy on two consecutive blocks). Whilst there is no suggestion of interleaving (as in, e.g., Lindsay & Gaskell, 2013), or of evoking the familiar competitor (as in, e.g., Kapnoula et al., 2015; Weighall et al., 2017), it may be that this over-training was responsible for strengthening the novel word sufficiently to allow engagement, in and of itself. It is however an open question whether training like that seen in the eye tracking literature would bring about immediate novel engagement in mouse tracking.

As a technique which appeared promising for later experimental work in this thesis, mouse tracking was examined further. The conclusions of that literature review are presented in Chapter 8.

CHAPTER

EIGHT

AN INTRODUCTION TO COMPUTER MOUSE TRACKING

8.1 What is computer mouse tracking?

Computer mouse tracking requires participants to view objects arranged on a screen, and click on them with a computer mouse (e.g., Freeman, Dale & Farmer, 2011; Freeman, 2018). As the mouse moves to the target object, various properties of the computer mouse path are measured (e.g., Freeman & Ambady, 2009; Kieslich & Henninger, 2017). Similarly to eye tracking, the logic of the technique is that by varying response options across experimental conditions, one may see the influence of those different response options on whatever processing is required by the task (e.g., Spivey et al., 2005). Mouse tracking tasks require discrimination between the response options, and allows participant decision processes to be imaged (Anderson & Spivey, 2009; Magnuson, 2005; Spivey & Dale, 2006).

A simple example of a mouse tracking experiment is seen in the paper introducing it. Spivey et al. (2005) sought to establish whether motor movements of the hand and arm, measured as a participant moved a computer mouse, could reflect lexical competition predicted by various speech perception models (e.g., the distributed cohort model (DCM); Gaskell & Marslen-Wilson, 1997). There were two experimental conditions: a target (e.g., CANDY) either appeared alongside a 'cohort competitor' (e.g., CANDLE), creating a 'phonological competition condition', or else alongside a distractor object (e.g., PICTURE), which competed only perceptually (and not phonologically). This condition was therefore the 'perceptual competition condition'. Critically, the distractor was not a phonological (onset) competitor and thus was not intended to interfere with responses to the target so much.

Seated in front of a computer screen, participants clicked on a button horizontally centred and at bottom of the screen to begin each trial. Five hundred milliseconds would then elapse, and pictures depicting the objects would appear on screen. Which pictures depended on the experimental condition as outlined above, and trials were interleaved, not blocked. At the same time as the pictures appearing, participants heard a word identifying the target object (e.g., 'candy'); they would then be required to click on the CANDY referent as quickly and accurately as possible. Objects

were rotated across conditions and appropriately counterbalanced to appear either as competitors, distractors or targets. The short delay between the trial-starting click and target appearance was to ensure that the decision itself could be imaged by the mouse path. Piloting showed that without this short delay, participants would sometimes wait for a word to be spoken, identify a target, and then move to it. The mouse path in this case would not index lexical competition *as it occurred*, but only the *result* of the decision process, as participants would move only once competition had been resolved. Instead, with the delay, participants would click, and then move to the objects, which always had predictable locations in the top left and right corners of the screen. Shortly after, they then heard a word, the onset of which either labelled both on-screen objects, or else only one. For example, the onset syllable /kænd—/ either labelled both the CANDY and CANDLE referents, or only the CANDY referent, but not the PICTURE referent. A typical set up is shown in Fig. 8.1 (p. 95).

But how does the mouse path index competition? Consider the comparison with tasks such as lexical decision (Meyer & Schvaneveldt, 1971) or pause detection (Mattys & Clark, 2002). In these tasks, participants' responses are slower and less accurate when a word to which they are responding has many competitors. However, when a word has no competitors, responding is maximally efficient (e.g., Bowers et al., 2005; Gaskell & Marslen-Wilson, 1997; Gaskell & Dumay, 2003). The same is true of mouse tracking: when the response options are PICTURE and CANDLE, and a participant has heard /k/, they may immediately respond (see filled circles in Fig. 8.1, p. 95). Likewise, the path of the mouse itself will be straighter. In this condition, competition is only *perceptual*.

By contrast, if a participant sees CANDLE and CANDY, they cannot respond until much later in the speech stream; here, not until the final /l/ or /iː/. Competition here is *phonological*. This will manifest as a longer response time (RT), and a mouse path which shows more attraction to the alternative response option, as it is a viable response for far longer (see unfilled circles in Fig. 8.1, p. 95). This inefficiency may be indexed in various ways – a review of the literature, presented later in this chapter, found that well in excess of ten measures have been used, each of which have different suitability, depending on the question being asked. Similarly, the hardware itself may be variable – in addition to using the standard computer mouse, work has also been done with a Nintendo Wii remote (e.g., Dale, Roche, Snyder & McCall, 2008; Duran, Dale & McNamara, 2010) or an electromagnetic track ball (Song & Nakayama, 2008).

8.1.1 Dynamic systems in mouse tracking

A useful conception of mouse tracking, and its responses, may be made if one considers a decision as being movement through multi-dimensional space – the 'decision landscape' (Fig. 8.2, p. 96). Several authors have conceived mouse tracking, and the selection of responses, as being akin to a marble rolling through this space, and with each response being a 'decision well', the depth and size of which is proportional to its attractiveness, according to the experimental condition (Magnuson, 2005; Spivey & Dale, 2006; Zgonnikov, Aleni, Piiroinen, O'Hora & di Bernardo, 2017). For ex-

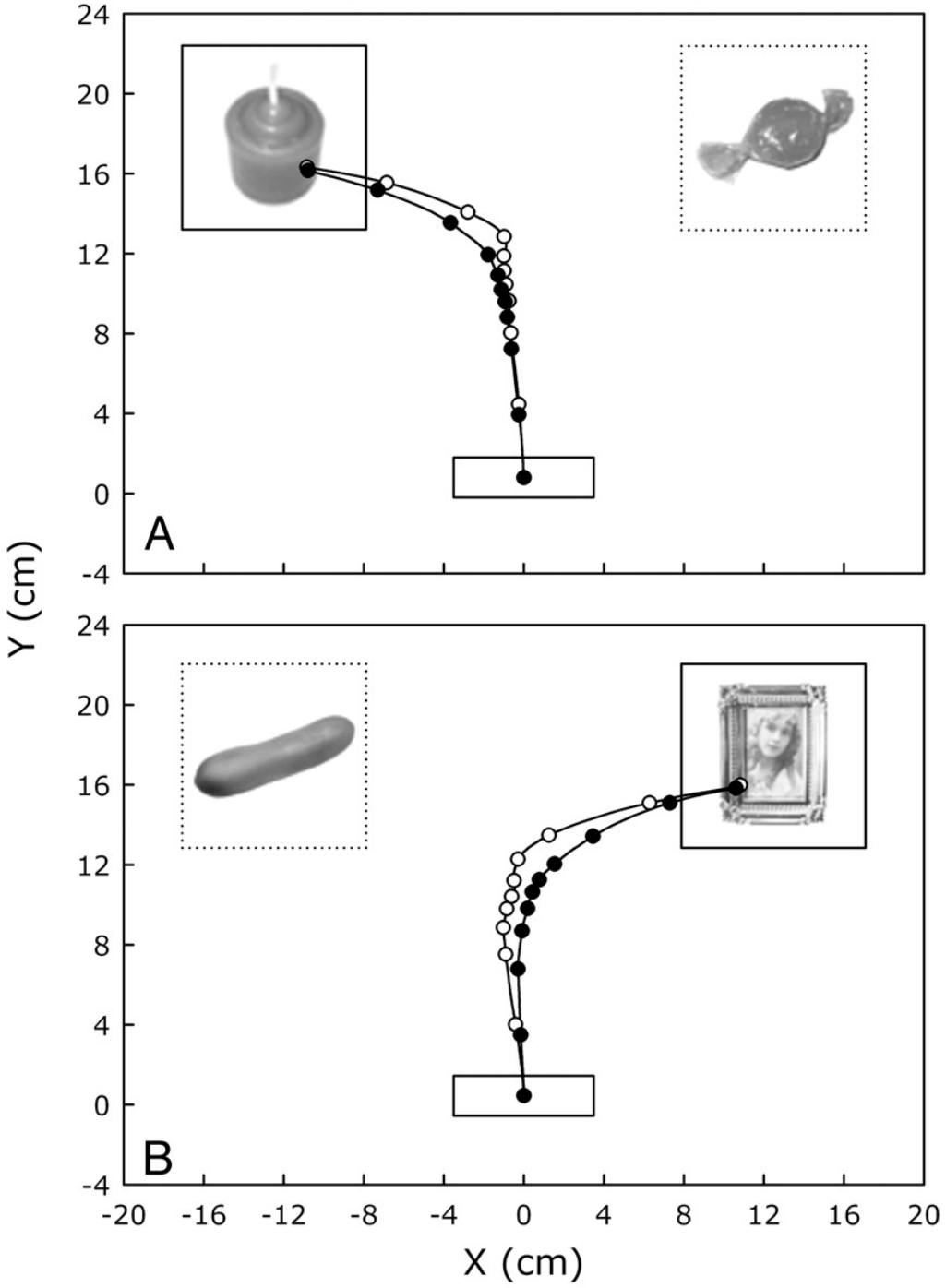


Figure 8.1: Leftward (A) and rightward (B) mouse trajectories shown plotted. Filled circles show the efficient responding in the perceptual competition condition, unfilled circles the inefficient responding in the phonological competition condition. However, in both images, phonological competition trials are shown: CANDY and CANDLE, and PICKLE and PICTURE. Adapted from Spivey et al. (2005)

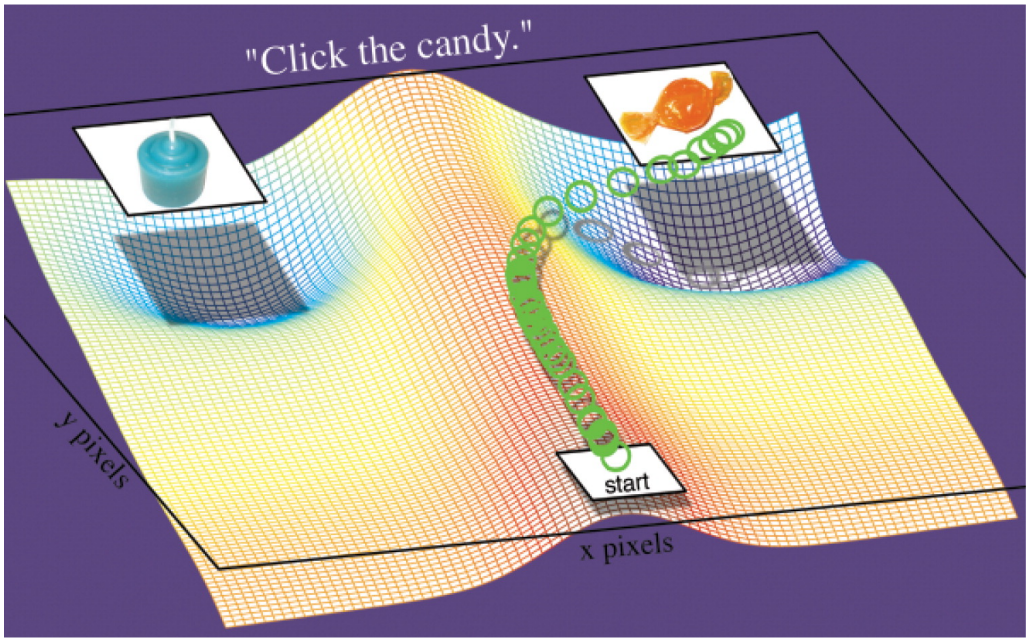


Figure 8.2: Mouse tracking conceptualised as movement through a 'decision landscape', and responses illustrated as 'decision wells'. Mouse path sampling illustrated by the green circles. Adapted from Spivey and Dale (2006)

ample, in Spivey et al.'s work, the 'well' of the target in the perceptual competition condition would be wide and deep, whereas the distractor would be shallow and narrow. However, in the phonological competition condition, the depth and breadth of the wells is more equal.

Spivey and Dale (2006) also advocate against a stage-based, modular view of decision making, as the motor response may be directed and influenced by linguistic inputs. Hence, the smooth and continuous movement, through a so-called decision landscape. The authors argue that although decisions may appear to be discrete, and for example, motor processing may appear to be subsequent and separate from language, evidence from mouse tracking shows that this is not the case (also, cf., Anderson & Spivey, 2009; Spivey, 2016). They argue instead that decisions are more analogous to a thread stitched through a hem – whilst each stitch of the thread may appear to be separate from its neighbour, closer inspection reveals them to be composed of one continuous thread. This framework should be kept in mind whilst considering the processing that takes place in a mouse tracking task, and it is worth noting that it is fits well with other psycholinguistic concepts. For example, the DCM models words as 'multi-dimensional arrays' (Gaskell & Marslen-Wilson, 1997). The model predicts that word recognition may be described by information on each dimension flowing into the system, causing it to move through space to a point of recognition. This conception of mouse tracking tasks and data is therefore compatible with the frameworks already set out in the initial chapters of this thesis (e.g., Chapters 1 and 2, pp. 3 and 7).

Table 8.1 Index of reviewed mouse tracking literature by research area

Research area	**Published work**
Consumer/marketing psychology	Johnson et al. (2012) Navalpakkam and Churchill (2012)
Language psychology	Barca and Pezzulo (2012, 2015) Barr and Seyfeddinipur (2010) Bartolotti and Marian (2012) Dale et al. (2007) Dale and Duran (2011) Farmer, Cargill et al. (2007) Farmer, Anderson and Spivey (2007) Spivey et al. (2005)
Mathematical cognition	Marghetis et al. (2014)
Memory	Dale et al. (2008)
Social psychology	Duran et al. (2010) Freeman and Ambady (2009, 2011) Hehman et al. (2014) van der Wel et al. (2014)
Vision research	Song and Nakayama (2008)

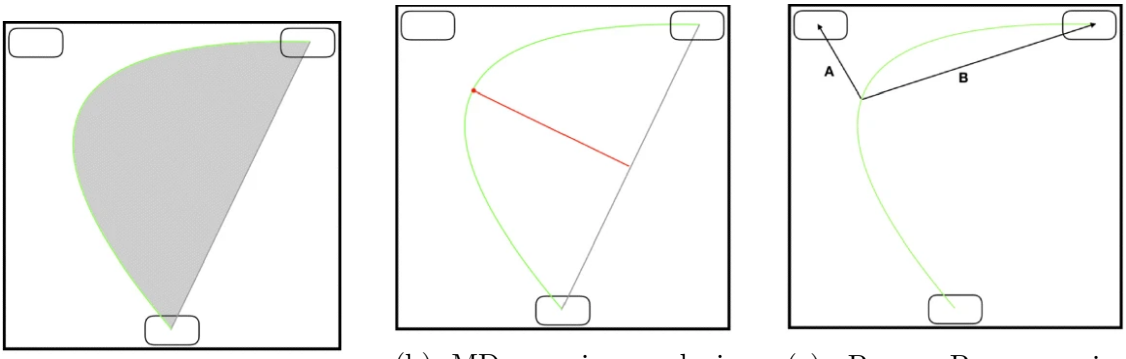
8.2 How has mouse tracking been used?

Mouse tracking has been used to answer a variety of experimental questions in a range of fields (Table 8.1, p. 97; for reviews, see Freeman et al., 2011; Freeman, 2018). Although not all of these papers are in the field of language, a review of the mouse tracking literature was undertaken to establish best practice for the technique, and to gain insight into which measures would be most optimal to report. Frequently, however, authors use a selection of the measures from those reported below – the data put out by mouse tracking is rich enough that different processes may be best represented by different measures. A summary of the measures follows.

8.2.1 The spatial dynamics of a mouse tracking response

As data is collected in real time, mouse tracking allows one to image responding as it occurs: for every participant, on every trial, the experimenter logs 60–75 mouse position data points per second (Freeman & Ambady, 2010; Kieslich & Henninger, 2017). This allows the experimenter to draw a decision path, which is then compared across conditions experimentally (e.g., Spivey et al., 2005). As the difference between response options is represented spatially on screen, clearly, the spatial component of the response is important.

The simplest way of analysing these data is to do so geometrically. The measures may be reported in a variety of ways: centimetres, pixels and arbitrary units



(a) AUC: area under curve

(b) MD: maximum deviation

(c) P_{min}; P_{max}: minimum/maximum position

Figure 8.3: Mouse tracking's spatial measures. AUC represented by the grey shading, MD by the red line. P_{min}/P_{max} shown by lines labelled A and B respectively, taken with respect to the competitor. Adapted from Maldonado et al. (2019)

(corresponding to a resolution-independent on-screen area) have all been used (e.g., Spivey et al., 2005; van der Wel et al., 2014). The literature reviewed identified four spatial measures: area under curve (AUC), maximum deviation (MD), the minimum/maximum spatial proximity to on-screen objects (P_{min}; P_{max}), and the length of the mouse path itself (PL). These are summarised below, and an illustration of these measures can be seen in Fig. 8.3 (p. 98).

- AUC: the area bound by the participant's trajectory and an idealised, straight line trajectory representing maximally efficient corresponding;
- MD: the point at which the participant's trajectory is furthest from the idealised trajectory;
- P_{min}/ P_{max}: the point of minimal/maximal distance from an on-screen object, usually the competitor;
- PL: the total distance travelled by the participant with the mouse cursor during responding (the curved green line in Figs. 8.3a to 8.3c).

Of these measures, the most popular is P_{min}/ P_{max}, which was used in six of the 19 papers reviewed. The least popular was MD, as it was only used in three papers. However, of the nine language papers reviewed, only one of them used P_{min}/ P_{max}. The most popular measure in language was PL, although still only three papers used it (Barr & Seyfeddinipur, 2010; Dale et al., 2007; Spivey et al., 2005).

8.2.2 The temporal dynamics of a mouse tracking response

The literature review identified four spatial measures, although there is some variability about how each measure is defined. They are initiation time (IT), movement duration (RT_m), response time (RT_t), and time spent in a region of interest (T_{ROI}). As T_{ROI} is used in a paradigm where the mouse is more freely allowed to move around the screen, instead of in the lexical competition paradigm described above,

it will not be discussed here. However, the remaining three are important, and measure important experimental parameters (e.g., Kieslich, Schoemann, Grage, Hepp & Scherbaum, 2020). However, temporal measures generally produce weaker effects than spatial measures in mouse tracking (Maldonado, Dunbar & Chemla, 2019).

IT measures the time at which movement begins after a trial is started (often with a mouse click, e.g., Spivey et al., 2005). In order to encourage movement, as described above, Spivey et al. (2005) introduced a 500ms stimulus-onset asynchrony. Later work uses, and experimental packages implement, an 'IT cut': a time point which, if exceeded, a trial is invalidated (Freeman & Ambady, 2010; Kieslich & Henninger, 2017; Kieslich, Schoemann et al., 2020). This increases the effect sizes observed (Kieslich, Schoemann et al., 2020), and often, researchers wish to demonstrate effects on other measures, in the absence of differences in IT (e.g., Spivey et al., 2005), as this would demonstrate qualitatively different processing across conditions. In a lexical competition task, this is clearly undesirable – different conditions should only pressure the system to greater or lesser degrees. However, where there are dual-system accounts of processing, and different conditions access these systems differently, some researchers have used IT to demonstrate theoretically interesting effects (e.g., van der Wel et al., 2014). Clearly, however, this is an important parameter. It was reported in seven of the papers studied, and three of the nine language papers.

The remaining two measures are not clearly distinguished in the literature. In theory, RT_t should equate to the total length of a trial, whereas RT_m should be equivalent to RT_t, less IT. Although this vocabulary is not always reliably used in the literature, given the pervasiveness of RT within cognitive psychology, its reporting in mouse tracking work is clearly a carry-over from older experimental paradigms (e.g., lexical decision; Meyer & Schvaneveldt, 1971), and allows for easy cross-trial and cross-experimental comparisons. All papers reporting RT_m also report RT_t, which is reported by four language papers, and a further three non-language papers.

8.2.3 Uniting spatial and temporal dynamics: trajectory analysis

In addition to being able to study either the spatial or temporal dynamics, the two measures may be united and studied together, as each change in position is accompanied by a time step. An advantage of mouse tracking is its very high temporal resolution: a recording can be made at least once every ≈ 16.7ms, $= 60$Hz, and more often on more capable hardware (Freeman & Ambady, 2010; Kieslich & Henninger, 2017). These type of measures are the most commonly reported in the literature, and 13 out of 19 papers reported some kind of trajectory analysis, including all the language papers. Analysis of standardised time bins in particular is very widely used in the literature (10 papers, 6 of which are in language psychology). However, there are four types of trajectory analysis in total:

1. **Standardised time bin analysis**. Instead of recording specifically how long a trial takes, trials may be carved into a number of equal sections (usually, 101: 100 time bins from start to finish of movement, plus the starting point), with no regard given to specifically how long any individual time bin is, on

a per-trial basis. Position in these aggregated bins is calculated by linear interpolation. Trajectories may have their start and end points aligned as well. These two procedures are referred to as time and space normalisation (Dale et al., 2007; Spivey et al., 2005). The analysis usually then compares conditions at a particular time bin, on either the overall trajectories, or possibly separately for the x- and y-vectors. The x- vector typically indicates response competition, whereas the y-vector typically indicates the overall attractiveness of both response options combined (e.g., Bartolotti & Marian, 2012). A comparison of x- or y-position at each time bin may be performed: 'runs' of time bins are then noted. Best practice is to only count runs of eight or more time bins, to control for the multiple comparisons (101, one for each bin; Dale et al., 2007). For example, for rightward trajectories, Spivey et al. (2005) found that there was a difference in attraction towards the competitor (i.e., a smaller x-position on phonological competition trials, indicating a more central position) for time bins 4–93. This revealed that lexical competition started very early indeed, and persisted for most of the trial. This matched well with theories of speech perception (e.g., Gaskell & Marslen-Wilson, 1997).

2. **Velocity profile analysis**. Fast movement is indicative of certainty: by tracking when the participant is moving fast or slow, one is able to infer the time course of cognitive events.

3. **Acceleration profile analysis**. An index of change in velocity, it indicates much the same thing.

4. **Trajectory type analysis**. Song and Nakayama (2008) compared curved and straight trajectories. However, as this research was outside the area of language, and not performed with a computer mouse, the measure will not be considered further.

8.2.4 Other measures: distribution analysis and decision dynamics

What follows is a brief discussion of measures not clearly related to the other dynamics. Broadly, these fit into two categories: analysis of the statistical distribution of responses, and analyses to extract some index of how certain or uncertain a participant is as they respond. However, unlike velocity and acceleration profiles, which also index response certainty, these certainty indexes are not firmly tied to the trajectories.

Distribution analysis: assessing non-unimodality

When examining the statistical distribution of the responses (for example, a histogram of the AUC by trial), one is most often seeking to avoid a multimodal distribution. A multimodal distribution would correspond to many 'types' of responses (see Fig. 8.4, p. 101). Given the aggregation that takes place to perform the trajectory analyses, these are problematic.

For example, consider the following scenario. A mouse tracking experiment is run whereby participants perform a word recognition task designed to measure lexical

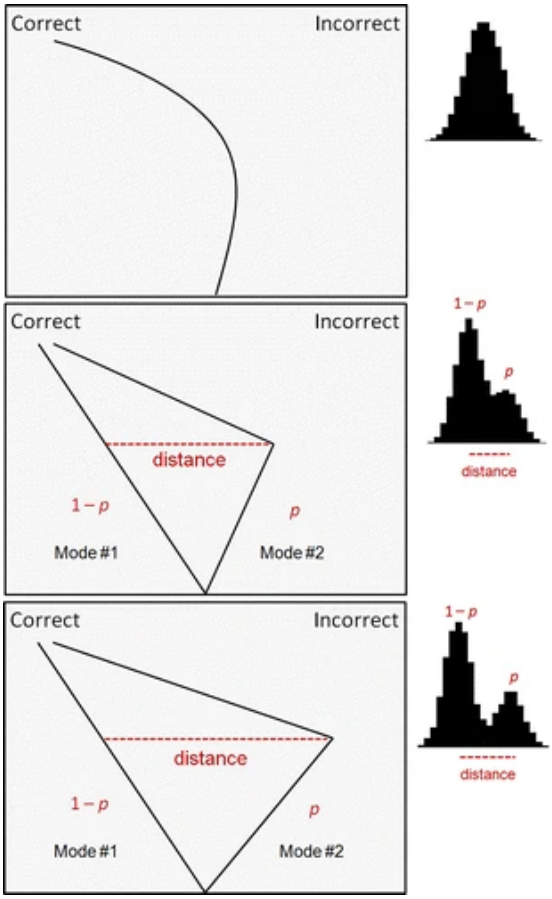


Figure 8.4: Illustration of uni- and bimodal responding. Note that all three panels have the same averaged curve, equivalent to that seen in the unimodal top panel. In this instance, histograms and distributional analysis are required to dissociate these quite different data patterns. Adapted from Freeman and Dale (2013)

competition, similar to Spivey et al. (2005). On the perceptual competition trials, all participants show no phonological competition and move efficiently towards the target. On the phonological trials, half of the participants again suffer little competition at all, and again respond with a straight line. The remaining participants also experience no *lexical* competition, but move rapidly towards one or other of the on-screen objects. Half the time, this half of the participants would be correct, but half the time, they would have to rapidly change course when they hear the target word, in order to select the correct object. This would create a sub-population of trials which are extreme – and not indicative of the smooth moving through a decision landscape that mouse tracking is supposed to measure (cf., Fig. 8.2, p. 96). More problematically, when averaged with the responders experiencing no lexical competition, there would be an *appearance* of lexical competition. This is because the small extreme sub-population would displace the trajectory in the phonological competition condition from the response exhibited in the perceptual competition condition.

One way to deal with this would be to inspect the data distribution visually. However, more sophisticated approaches exist. Freeman and Dale (2013) identify

three tests of non-unimodality: Akaike's information criterion between one and two-component distribution models (AIC_{diff}; Akaike, 1974), the bimodality coefficient (b; e.g., SAS Institute Inc., 2018), and Hartigans' dip statistic (HDS; Hartigan & Hartigan, 1985). Most commonly, the bimodality coefficient is used (e.g., Barca & Pezzulo, 2012; Dale et al., 2007; Farmer, Anderson & Spivey, 2007; Freeman & Dale, 2013; Spivey et al., 2005). It is calculated as follows:

$$b = \frac{S^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}} \tag{8.1}$$

where b is the bimodality coefficient, S is the skewness, k is the kurtosis and n is the number of observations. The bimodality coefficient considers a distribution bimodal if $b > 0.555$, as this the value for a uniform distribution. The bimodality coefficient is only capable of detecting if a distribution is bimodal or unimodal, and does not provide a significance value.

In their consideration of the two other bimodality measures, Freeman and Dale (2013) recognise that these have an advantage over the b as they are inferential tests. Of the two, Freeman and Dale's simulations showed an advantage for HDS, which was more sensitive than AIC, which was also less sensitive than b. Comparing b and HDS, Freeman and Dale found that the HDS is more sensitive, and less affected by skewness – b was found to report skewed unimodal distributions as bimodal. However, the authors state that the two often converge, and that all emphasise that b and HDS show clear advantages in sensitivity over AIC. This position has been supported elsewhere in the literature also (Pfister, Schwarz, Janczyk, Dale & Freeman, 2013).

Also considered by Freeman and Dale is the usual practice of removing outliers based on *SD*. The authors urge caution when doing this: as few as 5% of trials may be a detectable second mode, and they may be many *SD*s from the centre of the distribution.

Decision dynamics

Three other measures are used within the language psychology mouse tracking papers: the initial angle of movement, sample entropy (E_S), and x-position flips (XF).

1. **Initial movement angle** calculates, from the vertical, the angle at which a participant is moving a short duration into the trial. However, as the measure is not implemented in commonly used mouse tracking software packages, it shall not be considered further (Freeman & Ambady, 2010; Kieslich & Henninger, 2017).

2. **Sample entropy** (E_S) considers how stable the trajectory is over time (Dale et al., 2007; Hehman, Stolier & Freeman, 2015; Richman & Moorman, 2000). E_S compares 'windows' of the trajectory, with the size of the window specified. Likewise, another parameter defines a stability threshold. If the trajectory is not consistent across windows (with 'consistent' set by the stability threshold),

then it is entropic. High E_S would indicate highly disordered responding, implying more competition, as the trajectory changes unpredictably across windows. Typically, this is run on the x-vector of the trajectory, as this is the vector corresponding to competition between left and right placed responses.

3. ***x*-position flips** (XF) are a cruder, but more intuitive, way of measuring disorder, compared to E_S. Whilst E_S is sensitive to any kind of disorder – including that which may not necessarily result from a change in direction – XF simply count the number of directional changes. When a processing system is under stress and producing erratic responses, one would expect the number of XF to be high.

In addition to these measures, common to much of cognitive psychology, one may analyse the rate of error. However, mouse tracking tasks are easy enough that relatively few errors are made, even in higher cognitive load condition (e.g., Spivey et al., 2005). With such low error rates, analysis of these data is therefore spurious, particularly given the availability of more sophisticated measures unique to mouse tracking more directly measuring the decision dynamic.

8.3 The comparison of mouse tracking to eye tracking

Through the above, it is clear that there are many ways that mouse tracking data can be analysed, and indeed, that this is one of the methodological strengths of the technique. However, the same can be said of eye tracking, and as discussed in Chapter 7, eye tracking has much wider use in word learning research[1]. Moreover, eye tracking has been shown to produce pre-sleep lexical engagement, the effect of interest to this thesis. What then is the advantage of mouse tracking?

This issue is dealt with cleanly in the mouse tracking literature. Firstly, introducing the technique, Spivey et al. (2005) argues that mouse tracking has some clear advantages over eye tracking, although the techniques may be complementary (cf., Bartolotti & Marian, 2012). Whereas eye tracking results in ballistic, all-or-nothing responses – an object is either fixated, or it is not; a saccade is either launched, or it is not – mouse tracking shows truly continuous and smooth, graded responding. Whilst this may be approximated in eye tracking with fixation curves (i.e., the proportion of fixations to an object rising and falling over a time window, e.g., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Weighall et al., 2017), in mouse tracking, this is evident on every single trial. As the research questions in this thesis deal with the fragile representation of very newly learnt words, it is easy to see how this may be significant, and eye tracking may obscure an effect. This point is particularly important if one considers that these representations may

[1] Bartolotti and Marian (2012) is the exception to this – a mouse tracking paper studying novel word learning, *and* demonstrating pre-sleep competition effects. However, although the authors demonstrate lexical effects, this research is anomalous relative to the other literature, such as that work discussed in Chapters 3 and 7 (pp. 21 and 83). For example, the authors trained words in an explicitly L2 context, and required participants to select the novel word during responding – rather than measuring the extent to which a novel word may be activated by a known competitor (as in e.g., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Weighall et al., 2017).

only be partially activated and so cannot drive all-or-nothing eye tracking responses (Dale et al., 2007). As mouse tracking is a technique which allows one to measure the effect of "sub-threshold processes, [such] that deviations in smooth trajectories are observed even in the absence of visual saccades to a competitor" (Bartolotti & Marian, 2012, p. 1131), it is clearly preferable.

Secondly, although eye movements may be initiated *earlier* than skeletal movements of the hand/wrist/arm, it is not the case that eye tracking is *more sensitive*. For example, work studying if eye tracking may be replaced with mouse tracking for users interactions with web pages and adverts has found correspondence between eye tracking and mouse tracking (Johnson et al., 2012; Navalpakkam & Churchill, 2012). In a mouse tracking demonstration of immediate lexical engagement, Bartolotti and Marian (2012) found agreement between mouse and eye tracking in how monolinguals manage lexical competition. Moreover, Bartolotti and Marian leveraged the ability of mouse tracking to produce continuous data to detect a difference that was not present in the eye tracking data for bilinguals.

Supporting this point about sensitivity, mouse tracking produces many more data points than eye tracking: although eye trackers may *record* data at 1000Hz or so, processing and binning procedures may leave researchers with only a small number of fixations per second – effectively, reducing the data down to 5Hz or so (Spivey et al., 2005). This compares unfavourably with the 60–75Hz of mouse tracking (Freeman & Ambady, 2010; Kieslich & Henninger, 2017).

Thirdly, there is a practical point: the simplicity of mouse tracking makes it much more feasible. Whereas eye tracking requires particular and expensive equipment, and then particular skills to analyse the data, mouse tracking requires no specialised equipment, and although designs and analyses may be very sophisticated, they need not be. Likewise, there are no problematic issues around calibration of equipment etc., or the risk of data being unavailable due to factors such as blinking. Following the lack of a fast mapping effect in Experiments 1 and 2 (Chapters 5 and 6, pp. 55 and 65), this was an important consideration, as there was a need to quickly adapt this project to look for novel word lexical competition in other paradigms.

The final point is that eye tracking provided no clear model to follow in any case. Although there are four eye tracking papers in the literature demonstrating pre-sleep lexical competition, they do not follow one particular procedure. Bartolotti and Marian (2012) examined how lexical competition, with highly over-trained novel words, may be managed in a second language by either mono- or bilingual speakers. Those authors' task required participants to click on the *novel* object – whereas responses are more typically to a known object, without further activation of the novel object. Kapnoula et al. (2015) and Kapnoula and McMurray (2016a) used an auditory mismatch paradigm, paired with the visual word paradigm (VWP; Tanenhaus et al., 1995). However, this protocol does not allow one to look at semantics, as Kapnoula and colleagues' effects are based solely on phonology.

Weighall et al. (2017), however, *did* train semantics, and did use a procedure more in line with the papers from the literature review (Chapter 3, p. 21), asking "What affect does learning '*biscal*' have on the processing of 'biscuit'?" (cf., Gaskell & Dumay, 2003). However, there were some surprising findings in that paper, and given the above, mouse tracking had certain advantages in studying novel word

learning. For example, no difference was shown between words learnt on the day of, and the day before, testing, in contrast to other research (e.g., Dumay & Gaskell, 2007). Whilst the design fit well with the rest of the literature, and the effects were interesting, use of eye tracking may have obscured some novel word effects, according to arguments set out in the mouse tracking literature. A direct methodological eye tracking replication may therefore have been sub-optimal. Thus, mouse tracking was chosen for future experiments.

8.4 Conclusions: how may mouse tracking be applied to word learning?

As mouse tracking had been settled upon as the technique of choice for future work, the next step was how to implement it. Given the above, several questions stood out.

1. Could the simple design of Spivey et al. (2005) be easily and quickly implemented and replicated, with a laboratory set up and data analysis procedures to detect phonological competition?
    (a) This would involve establishing some simple experimental parameters, such as the position and size of objects on screen, as this is variable in the literature (usually in the range 200 × 200 to 400 × 400 pixels)
2. For each of the 'categories' of measurements identified above, which was the most appropriate for language research?
    (a) Indeed, were all the categories necessary? For example, would disorder analyses reveal anything beyond that shown by the temporal or spatial dynamics?
3. Was mouse tracking robust to changes in stimuli type, or design?
    (a) It was noted that whereas Spivey et al. rotated items across conditions, the novel word paper which appeared to be the best candidate for follow-up study and replication, Weighall et al. (2017), did not use this design. Instead, that work presented each item only once. Moreover, whereas Spivey et al. used larger, photo-realistic stimuli, Weighall et al. used smaller cartoon stimuli. As the novel words were liable to produce weak effects (perhaps due to their being on partially activated during responding), it was considered that pilot work with known words may first be needed to investigate just how relevant these two changes would be.

With these questions in mind, two pilot experiments were undertaken (Study 2). Firstly, Experiment 3 (Chapter 9, p. 107) undertook a replication of Spivey et al. (2005), in order to establish basic experimental and data analysis protocols. Secondly, Experiment 4 (Chapter 10, p. 129) made the changes to the design to test the robustness of the effects which were observed. This work was then fed forward to novel word studies in Experiments 5, 6 and 7 (Chapters 12 to 14, pp. 155, 177 and 191).

CHAPTER

NINE

EXPERIMENT 3
MOUSE TRACKING LEXICAL COMPETITION

9.1 Introduction and rationale

Chapter 8 concluded that mouse tracking was a technique that would be ideal for studying novel word learning. However, before attempting a series of novel word learning experiments, it was important to demonstrate, and develop a set up to detect, phonological competition in familiar words. This meant attempting a simple mouse tracking pilot experiment, in order to gain the required skills for designing and running such projects, and analysing the resultant data.

To pilot mouse tracking, a seminal paper was targeted for replication: Spivey et al. (2005). In that experiment, participants chose – by mouse click – one of two on-screen objects in response to a stimulus. Here, the stimulus was a single spoken word (e.g., 'dolphin'). On screen would be a target, and either a phonological onset competitor (a 'cohort competitor'; cf., Gaskell & Marslen-Wilson, 1997) or a distractor. The target object was a photograph of the word's referent (i.e., DOLPHIN), and the distractor/competitor was also a referent of a concrete noun which either did or did not overlap with the heard word (e.g., DOLLAR or GUITAR).

Data were compared across the two conditions within-subjects. The control condition, with the distractor, only required participants to identify the correct referent for the heard word. For example, a participant would hear 'dolphin' and be required to select DOLPHIN and reject GUITAR. Responding here only required resolving *perceptual* competition between objects, and established a baseline level of response competition experienced by a given participant between pairs of objects.

This was contrasted with a test condition, whose objects had labels which also competed *phonologically* (e.g., DOLLAR and DOLPHIN). With the additional cognitive demand of rejecting a similar-sounding object label, Spivey et al. found participants responses became more inefficient across a series of measures: the area under the curve of the mouse path (AUC) increased; the mouse moved closer to the competitor (and further from the target); the length of the mouse path increased; and the response was slower (both with and without the initial time to initiate a movement counted) and more error-prone. However, this occurred in the absence of differences

in initiation time (IT) – indicating that the effects were due to competition *during* responding, and that participants were not selectively resolving competition on perceptual trials, and only then moving to a target. Supporting this argument, the distribution of the z-scored trajectory curvatures was found to be unimodal, with all participants responding in a similar mode. Finally, analysis of time and space normalised trajectories in the x direction revealed significant differences for up to 83 of the 101 time slices.

9.1.1 The present experiment

In the 15 years since Spivey and colleagues ran their study, mouse tracking has advanced quite a lot (e.g., Calcagnì, Lombardi & Sulpizio, 2017; Freeman, 2018; Hehman et al., 2015; Zgonnikov et al., 2017), and a wide variety of measures, hardware, software, and experimental protocols have been used (see Table 8.1, p. 97). Although readily available software packages and published work on designing mouse tracking experiments made identifying appropriate design options easier, the recommendations are necessarily generic, and there is acknowledgement that researchers should identify locally what works best for their experimental questions (Freeman & Ambady, 2010; Kieslich & Henninger, 2017; Kieslich, Schoemann et al., 2020). A central aim of Experiment 3 was therefore to test out some design choices that had been made. These choices would then be regarded as successful in the event of replication of Spivey et al.'s key findings.

Design parameters and differences from Spivey et al. (2005)

The labelling task. One potential problem was that multiple forms could map to the same referent. For example, a picture of a bank note may plausibly also take the forms 'bill', 'note', 'fiver', and so on. To counteract this, a labelling task was introduced, to train participants on the intended forms. It should be noted that although the labelling task was a deviation from Spivey et al.'s original design, it does have precedent elsewhere in the literature (Kapnoula et al., 2015; Kapnoula & McMurray, 2016a).

The bimodality problem. Within mouse tracking methodology, non-unimodality has been identified as a theoretically important aspect of the distribution of trials (e.g., Hehman et al., 2015; Freeman & Dale, 2013; Spivey et al., 2005), as it allows one to distinguish between one-stage and two-stage accounts of processing. As a test of lexical competition, predicted by speech perception models (e.g., Gaskell & Marslen-Wilson, 1997), processing here should have been unimodal. A unimodal distribution of trials would imply a consistent mode of responding across all trials, and has been found in previous work (e.g., Spivey et al., 2005). By contrast, as illustrated in Fig. 8.4 (p. 101), a bimodal distribution of trials would imply two separate modes of response. In one mode, a participant moves straight to a target, without experiencing competition. In the second mode, a participant shows some movement to the competitor, which is then corrected. This is problematic, as it does not typify 'graded, continuous' responses of a participant moving towards a target,

but only *appears* to do so in aggregate. It was hoped that the data in Experiment 3 would be unimodal, as without unimodal data, analyses could not be performed in the way described in Section 8.2 (p. 97).

Speed. In mouse tracking, an important parameter is mouse 'speed'. This is a metric which (non-linearly) scales the 1000 DPI movement resolution of the mouse to determine on-screen pixel movement relative to the real-world movement. A mouse set to high 'speed' would move a long way on the screen with only a small real-world movement. As the experiment sought to capture physical movement, and movement at the default speed would result in more ballistic responding (i.e., a small skeletal movement resulting in a large on-screen movement, which would then need a secondary movement to correct, e.g., Sandfeld & Jensen, 2005), it was imperative that the mouse speed was slowed as much as possible, whilst at the same time leaving the speed fast enough so that responding was smooth – and thus continuous – with no jerkiness in the movements (cf., Kieslich, Schoemann et al., 2020).

Stimuli. Instead of using Spivey et al.'s stimuli, stimuli were created locally. This allowed participants to hear the words in a familiar accent, and ruled out any American English items. The removal of American English items was done to allow for phonological competition on trials which would have been perceptual competition trials in British English. For instance, consider the pair PICTURE – PICKLE. As participants were speakers of British English, it was thought that they would have instead mapped the form 'gherkin' to the PICKLE, forbidding phonological competition with 'picture'.

In creating the stimuli, it was noticed that the recordings were longer than those used by Spivey et al. (M = 532ms). However, it was assumed that this was due to differences between the speakers (e.g., accent, talking speed) and different items being used, rather than differences in how the stimuli were produced, or how well they were cropped.

Although the images used in Experiment 3 were not normed, care was taken to select images which were stereotypical and without any unusual or particularly salient features or colours (see Fig. B.1, p. 215). The words to which the images referred had all previously been used in published research or else were common objects (Spivey et al., 2005; Weighall et al., 2017, see Table B.1, p. 215, for a full list of experimental stimuli). Objects were selected for their being concrete nouns that one could reasonably assume to be familiar to the participants. All words were two syllables in length.

Trial numbers. To give more data, and it was hoped, stronger effects, the number of trials was increased compared to the work of Spivey et al.. Participants contributed more data points per item to balance out any unusual responses and increase statistical power. Whereas Spivey et al. only took data from 32 trials, here participants completed 96 trials.

9.2 Methods

9.2.1 Participants

Data were taken from 38 fluent English speakers (16 male, $M_{\text{age}} = 31.5$ years, $SD_{\text{age}} = 18.8$ years, 32 monolingual, 36 right handed). All were free of any learning and language disorders, and had normal or corrected to normal hearing and eye sight. All were right handed mouse users, as recommended in the literature (Kieslich, Schoemann et al., 2020).

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, or in exchange for course credits.

9.2.2 Materials and apparatus

The experiment was conducted in MouseTracker (Freeman & Ambady, 2010), on a 60 Hz, 19″ monitor, with a display resolution of 1280 × 1024 pixels. A commonly available USB laser mouse, the Logitech RX250, was used for data collection, polled at a minimum of 60Hz (Freeman & Ambady, 2010). The mouse had a movement resolution of 1000 DPI (dots per inch), and a speed multiplier of 1.75 was selected. This was in line with recent work (e.g., Feather, Vélez & Saxe, 2014). Words heard by the participants were delivered in a quiet laboratory environment over good quality headphones.

The words delivered to the participants were recorded on a single 44.1kHz channel, sampled at 32 bits, in a male voice, with a northern English accent. Recordings took place with a good quality microphone in a sound attenuating booth, to minimise unwanted noise on the recordings. These were cropped closely, so that only a single object label was heard, without significant onset or offset in the speech signal, and the amplitudes were normalised in Praat to 60dB (Boersma, 2001). The mean length of the recording of each word after cropping was 732ms ($SD = 183$ms). Differences between mean recording length was not significantly different across the three word categories used (Kruskal-Wallis test performed due to non-normal sample, $H(2) = 0.945$, $p = 0.623$).

The recordings were presented in MouseTracker with images found by Google Image search. The images were edited to remove any background, centred, and scaled to 300 × 300 pixels. As the pictures themselves formed response boxes for the participants' mouse clicks, when they were presented in MouseTracker the pictures were made to appear with a thin black border[1] surrounding them, so that participants could easily differentiate between what was and was not a valid place to click.

Additionally, a further 24 images were chosen to act as distractors, for use in a labelling task (see below). As these objects were never named, there was no associated recording. However, it was assumed that participants would recognise these objects, as they were mundane and commonplace (e.g., TABLE). These objects

[1]The MouseTracker documentation does not specify, or allow configuration of, the point width of this border. However, the border was not noticeably or distractingly thick, in order to allow the images to be as salient as possible.

Table 9.1 An overview of the design in Experiment 3, showing the rotation of items from a single triplet across experimental and filler trials. The full set of items can be seen in Table B.1 (p. 215)

Experimental trials		**Filler trials**
Perceptual competition	*Phonological competition*	
DOLPHIN × GUITAR	DOLPHIN × DOLLAR	GUITAR × DOLPHIN
DOLLAR × GUITAR	DOLLAR × DOLPHIN	GUITAR × DOLLAR

Note. The target of the trial is indicated by the red font. Spatial arrangement of items on screen during testing is not indicated here: the target appeared on both the left and the right.

were neither semantic nor phonological competitors to any of the other objects they were displayed with (that is to say, they shared no onset phonemes, and each was from a different class of objects). These objects were seen only in a labelling task, not in the experiment itself.

9.2.3 Design

Using a within-subjects design, the experiment took 24 word-picture pairings and arranged them into eight sets of three pairings. The words from the first pairing and second pairing of each triplet overlapped with each other on their first syllable: for instance, 'dolphin' and 'dollar' sharing the initial syllable /dɐl—/. As participants only heard the word for one of the pictures per trial, which picture was named alternated between trials, with all pictures appearing as both a labelled target and as a competitor. The allocation of items to be either the first or second in a triplet was arbitrary and irrelevant.

The final member of the triplet was an item which did not phonologically compete with either of the first two items, as its label differed from the first phoneme. The ordering of the word-object pairings into triplets remained fixed across the experiment, and every item only ever occurred with the other items from its triplet (see Table B.1, p. 215).

An example set of trials is set out in Table 9.1 (p. 111), where the object printed in red had its label pronounced, and was thus the target. Target-left and target-right presentation was counterbalanced with a single presentation of each trial type shown in Table 9.1. This gave a total of 12 conditions once left and right presentations were accounted for (6 from Table 9.1, × 2 for right and left). Therefore, with eight triplets, this produced $8 \times 12 = 96$ test trials, of which a maximum of 64 were analysed per participant, and 32 were discarded as filler trials.

Experimental trials were of two types, as in Spivey et al. (2005). Trials where the target appeared alongside a cohort competitor were *phonological competition trials*. Trials where the target appeared alongside a distractor object were *perceptual competition trials*. This gave the only independent variable manipulated in the experiment: Competition, with the two levels *Perceptual* and *Phonological*.

Filler trials, where the target was the distractor object, were included to ensure attention was not biased against perceptual competitors. This was a concern, as distractor objects were not otherwise the targets of trials.

In addition to the 96 test trials, there were a further 24 labelling task trials (one per word-object pairing, discussed below). Therefore, participants completed 120 trials in total.

9.2.4 Procedure

The labelling task

The experiment had two phases, which ran in a fixed order. First, participants completed a labelling task, trials of which also acted as a practice for the participants (consistent with suggestions by Kieslich, Schoemann et al., 2020). The purpose of the labelling task was to allow participants to learn which phonological representation each image was designed to trigger. For instance, that the picture DOLLAR was supposed to be associated with 'dollar' and not 'money', 'bank note', 'cash', etc. This training was necessary in order to ensure the *possibility* of phonological competition, as of the possible labels, only one (i.e., 'dollar') would compete in the intended way with its competitor (i.e., 'dolphin'). Secondly, it also allowed participants to become accustomed to the slower-than-default mouse speed.

In the labelling task, participants saw a set of instructions telling them that pictures would appear in the top left and right corners of the screen, and that they would hear a word for one of these objects. They were told that their task was to click on the object for which they heard a word as quickly and accurately as possible. They were also told that the mouse may feel different to them, but not to be perturbed by this, and to try to move the mouse as smoothly as possible, without picking it up, or grinding the mouse against the desk in attempt to get more traction or make the mouse move faster.

Participants started each trial by clicking a button labelled START, centred around the origin. When this button was clicked, participants' mouse cursor position was set to the origin `(0,0)`, no matter where on the rectangular button they clicked. Simultaneously, two pictures (300 × 300 pixels in size) appeared, placed equidistantly from the vertical edges and the midline of the screen (170 pixels on a 1280 × 1024 display), and 100 pixels from the top of the screen. Each picture formed a response box, clearly defined by a black border line. Participants were instructed that the pictures would appear in the top left and top right quadrants of the screen, and that following their click to start the trial, they should begin moving towards the top of the screen and the images. After 500ms of silence, a word labelling one of the pictures was heard, and participants then had to click on the referent picture. For instance, when a participant heard the word 'dollar', they were expected to click on the DOLLAR, ignoring either the DOLPHIN or GUITAR, depending on the condition (see Table 9.1, p. 111). Trials ended when a click was registered inside the response box drawn around the target referent. The START button then re-appeared after a further 500ms, which participants clicked to begin the next trial. Progression was therefore self paced.

Trials in the labelling task featured the 24 items for the experimental trials, each

pictured against a commonplace object (assumed to be familiar to the participants), such as a pair of headphones (for details, see Table B.1, p. 215). These 'labelling items' were never named, as the focus of the labelling task was for participants to learn the intended labels for the experimental items. All trials appeared in a random order.

The experimental task

Once they had completed the labelling task, participants proceeded to the experiment proper, which featured only the experimental and filler items (Table B.1, p. 215). Unlike in the labelling task, all objects appeared both as target and as perceptual/phonological competitor (Table 9.1, p. 111). Participants were offered another opportunity to do the labelling task if they felt they needed more practice; none stated that they did.

The experimental block of trials was performed in a very similar way to the labelling task. Participants saw the instructions again, reminding them to move the mouse smoothly, and to click on their target as quickly and as accurately as possible. Again, each trial started with a click of the START button, and participants began moving soon afterwards. After 500ms, a label for one of the on-screen objects was played. Trials ended with a click in the response box for one of the objects. After a further 500ms, the START button would reappear to begin the next trial. Participants were able to complete the whole experiment in under 15 minutes, and were thanked for their time at the end.

Measures in Experiment 3

A central question of Experiment 3 was which measures produced the strongest effect sizes on this particular psycholinguistic task. Strong effect sizes were favoured in order to offset any potentially fragile effects from the novel word representations studied in later experiments. As a first experiment, it was decided the best approach was to exclude no *type* of measure. Therefore, the following measures were analysed:

- Distributional analyses:
  - Freeman and Dale (2013) concluded that whilst Akaike's information criterion (AIC) was insensitive, the commonly used bimodality coefficient (b, Eq. (8.1), p. 102; SAS Institute Inc., 2018) and Hartigans' dip statistic (HDS) were often in agreement, although b was more biased by skewness. Therefore, both b and HDS were reported.
- Decision dynamics:
  - As anticipated, errors were few in number, and their analysis would not have been meaningful, and so was not conducted (see Section 8.2.4, p. 103). However, for reference, the number of errors made was reported.
  - Of the two other viable measures, x-position flips (XF) and sample entropy (E_S), both were reported, to compare effect sizes relative to each other, and the spatial and temporal measures.

- Spatial dynamics:
  - Area under curve (AUC), maximum deviation (MD) and path length (PL) were all viable measures, and so all were reported to compare effect sizes. The maximal/minimal proximity to the competitor (P_{min}/P_{max}) was dropped, as the point to calculate it from was not apparent with large images.
    - Given the size of images, the strength of the measure would have depended upon this design choice, and not the participants' responding, making any reporting of the measure spurious.
- Temporal dynamics:
  - Initiation time (IT) was emphasised in the literature as an important 'check' measure, and so was reported. Response time (RT) was then calculated and reported, due to its prevalence in cognitive psychology, and as a check of the spatial measures (i.e., as larger spatial measures should have implied longer RTs). The RT value reported here was the total amount of movement time, less participant IT, calculated on a per-trial basis.
- Trajectory analysis:
  - Whilst it was felt that velocity and acceleration profiles were difficult to interpret, especially as certainty could be more easily established by the decision dynamics metrics, separate x-vector standardised time bin analyses were quintessentially why one would want to use mouse tracking. However, due to unexpectedly bimodal data, this analysis could not be reported (see Table 9.3, p. 120).

In summary, therefore, eight analyses were performed on the data. It was recognised that this was a high number (the highest used in the literature being eight; Bartolotti & Marian, 2012; Spivey et al., 2005); however, it was warranted, as in this first study, a central aim was to establish the usefulness, strength and viability of different mouse tracking data components. It was intended that future experiments would use many fewer measures.

Processing of the mouse tracking data

Analysing the mouse tracking data required analysis of two data sets: participant trajectories, and then the derived measures calculated by contrasting these trajectories with idealised mouse paths. Various manipulations were performed on the data before it was subjected to inferential testing:

- All trajectories were remapped to the right side of the screen, pooling left and right presented targets, by reflection of leftward trajectories along the y-axis. Trajectories were also transformed so that they all ended at the same point of `(1,1)`. This controlled for participants clicking in different parts of the

response box, and is referred to as 'space normalisation' (Dale et al., 2007; Spivey et al., 2005). Note whilst some authors have analysed left-side and right-side target trials separately (e.g., Spivey et al., 2005), it was not deemed a variable of interest in this case, and trimming procedures meant all trials were within $3SD$ of the means, regardless. Furthermore, certain authors advocate collapsing across left and right sided responses to maximise statistical power (e.g., Dale et al., 2007).

- If participants did not immediately move from the origin at the trial start, this period of inactivity was removed from the RT recorded. All RTs were therefore reflective of the time that a participant was moving until the click in the response box. This was performed on a per-trial basis.

- As the rate at which the mouse was sampled by the computer was somewhat variable ($M = 15.9$ms, $SD = 4.26$ms, IQR = 11–20ms, maximum = 69ms) due to hardware limitations, the sampling was standardised to 20ms steps. The data for any resampled time points between recorded steps were filled in by linear interpolation. This was performed per trial.

- Again on a per-trial basis, trajectory data points were then put into 101 time bins, and the data values again (linearly) interpolated. This is the 'time normalisation' procedure conducted by Spivey et al. (2005), and controlled for the fact that the trials were subject to variable RTs, e.g., due to variable stimuli lengths.

Exclusions

Trials/participants were excluded for the following reasons:

- Filler/labelling task trials;
  - It had been the intention to reject participants based on labelling task performance, however, all participants performed at 100% accuracy.

- Trials with an incorrect response (56 trials; 1.54% of all trials);
  - All errors were for phonological competition trials. Low error rates confirmed the decision made above not to analyse error trials, due to their infrequency.

- Any trial with a value exceeding $M \pm 3SD$;
  - For each measure, a condition mean was calculated, and trials were cut off below/above $3SD$ from the mean. The trial as a whole was removed – so even if, for example, the XF were within the $\pm$ $3SD$ range, but outside it on RT, the trial was still removed. The mean was calculated across participants.

  - Although Freeman and Dale (2013) urged caution when applying a SD cut, visual inspection of the trajectory data showed that some participants had responded strangely, for example, by looping. It was felt it was more important to remove these irregular trials, and the only obvious way to do so efficiently was with trimming. It was assumed that rather than being a genuine second 'mode' of responding, this essentially amounted to random noise.
  - This removed a further 167 trials.

- If a participant had $< 75\%$ of their trials remaining;
  - This resulted in the exclusion of five participants and a further 228 trials.

The final data set therefore contained 1981 trials (997 perceptual, 984 phonological) from a total of 33 participants, meaning on average remaining participants contributed 60.0 trials (out of a maximum of 64). Most excluded trials were removed due to trimming procedures, and during the removal of participants entirely, not due to error. Low error rate meant that it was less likely that there were technical problems with the experimental parameters or procedures, or issues with participants understanding the task. Trimming procedures were reasonable to perform, as they ensured that the data remaining came from a sample responding in broadly similar terms. Remaining participants contributed 93.8% of their possible trials. This implied that the number of trials removed were not evenly distributed across the sample, but almost exclusively from particularly poor participants, who were then rejected when they were found to be contributing a low number of trials.

9.3 Results

Mouse tracking data were collected in MouseTracker (Freeman & Ambady, 2010). All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016), and the mouse tracking data were processed with `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff, Henninger, Haslbeck & Brockhaus, 2020).

9.3.1 Descriptive statistics

Table 9.2 (p. 117) displays the descriptive statistics for each type of competition. Participants suffered both spatial and temporal disruption when responding to a pair of objects with phonologically competing labels (e.g., 'dolphin' and 'dollar'), relative to objects whose labels only competed perceptually (e.g., 'dolphin' and 'guitar'). This did not appear to be due to different ITs. Participants' decision making also appeared more laboured for phonological competition trials than for perceptual competition trials: they were less likely to maintain a stable trajectory over time (as evidenced by E_S), and more likely to change direction (as evidenced by XF).

Table 9.2 Summary of descriptive statistics per level of *Competition* for each measure in Experiment 3

Measures	Competition			
	Perceptual		**Phonological**	
	M	*SD*	*M*	*SD*
Area under curve (AUC)	0.173	0.218	0.306	0.151
Initiation time (ms, IT)	402	318	414	334
Maximum deviation (MD)	0.320	0.274	0.568	0.242
Mouse path length (PL)	1.77	0.273	2.25	0.336
Response time (ms, RT)	1295	348	1428	345
Sample entropy (E_S)	0.074	0.020	0.092	0.023
x-*position flips (XF)*	1.15	0.709	1.45	0.679

Note. The units for AUC, MD, and PL are arbitrary.

Correlations between mouse tracking measures

An aim of this experiment was to find the most appropriate mouse tracking measures, and identify which measure was best suited to capturing a particular dynamic. Whilst effect sizes would show which measure was strongest for a particular dynamic, correlations would show the degree to which the measure measured the same thing, and thus how valid it was to compare effect sizes. This analysis was performed for the spatial and decision dynamic measures, but not for the temporal dynamic measures, as they were derivatives of one another[2].

Correlations between measures were examined in blocks per competition type. The spatial measures showed particularly strong positive correlations (see Figs. 9.1a to 9.1c, p. 119). The decision dynamic measures showed positive correlation, but a little weaker (Fig. 9.1d, p. 119). This implied that within a dynamic, the proposed method of comparing measures by their effect size and selecting the strongest was valid.

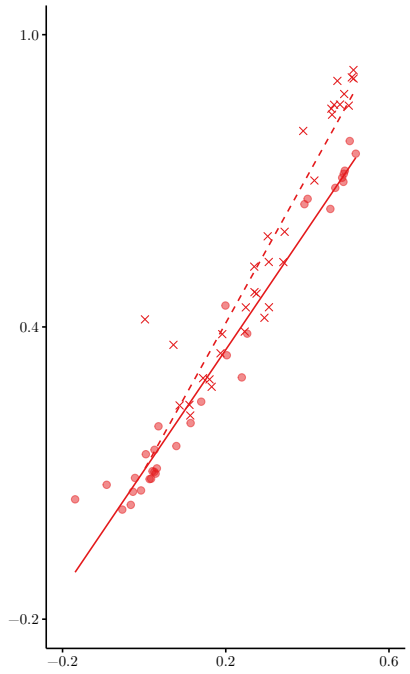
Distributional analyses

Bimodality was reported for two measures: the bimodality coefficient (see Eq. (8.1), p. 102), and Hartigans' dip statistic (HDS). HDS is interpreted as any other inferential test is: a p value of ≤ 0.05 indicates a departure from unimodality (HDS returns $p \leq 0.05$ where there is a sizeable inflection in the distribution). The bimodality coefficient (b; SAS Institute Inc., 2018), is interpreted against a threshold of 0.555, which is the value it returns for a uniform distribution (smaller values indicating a 'pinching' of the middle of the distribution – indicating unimodality, and greater values indicating a 'pulling up' of either end of the distribution – indicating bimodality). Table 9.3 (p. 120) shows the bimodality statistics for the data in Experiment 3. Fig. 9.2 (p. 120) shows the accompanying histograms. As multimodality is only a concern for the spatial measures, bimodality statistics were only calculated for the three measures AUC, MD and PL.

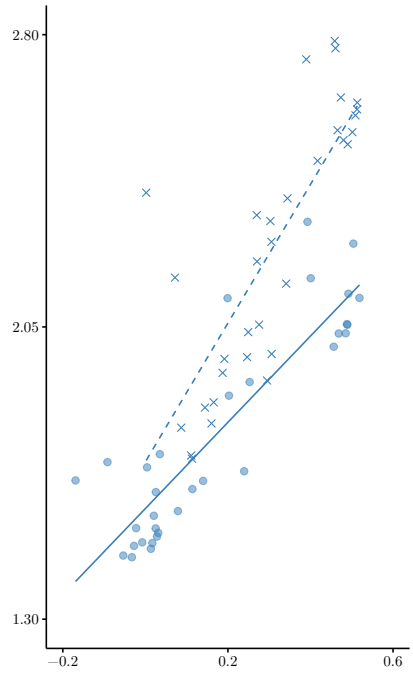
In accordance with Freeman and Dale (2013), the two bimodality measures were examined for convergence. Convergence was observed for all measures, except for the AUC phonological competition data. Here, b reported a low statistic, indicating unimodality, whereas HDS reported that the distribution was very unlikely to be unimodal. Visual inspection of the histograms (Fig. 9.2, p. 120) confirmed that HDS seemed to be the more reliable statistic: all the distributions showed clear second peaks.

Further explorations of bimodality. The finding of bimodality was unexpected, and represented a failure to replicate Spivey et al. (2005). Initially, it was assumed that this was caused by two sub-samples each responding unimodally. One characteristic of the sample in Experiment 3, relative to previous experiments involving mainly undergraduates, is that there were more older people (reflected in the larger mean age, and wider age standard deviation). This occurred because testing took place in two phases, one of which was over summer, when undergraduates

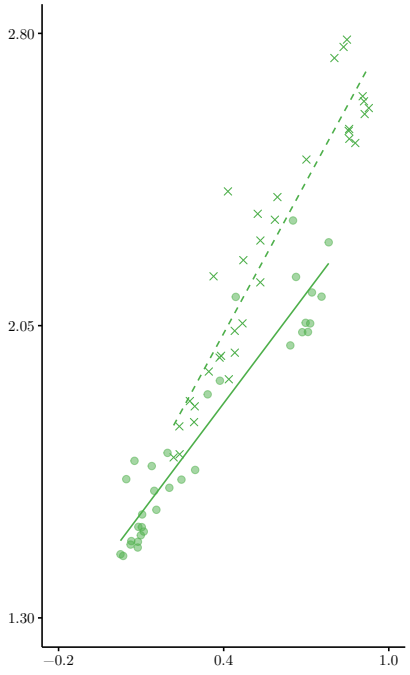
[2]As the total trial length was equivalent to IT + RT



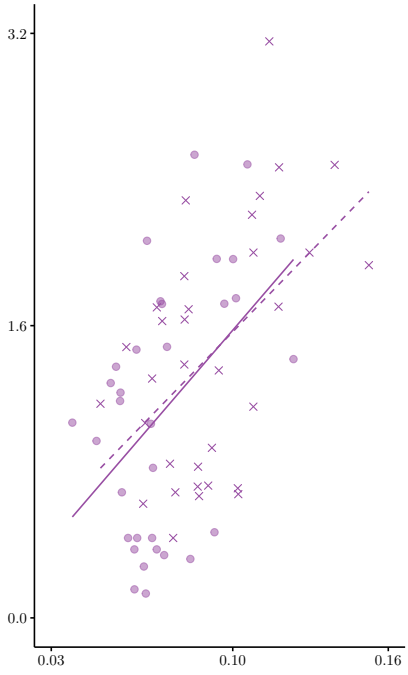
(a) AUC × MD ($rs = 0.983, 0.939$)



(b) AUC × PL ($rs = 0.882, 0.799$)



(c) MD × PL ($rs = 0.944, 0.938$)



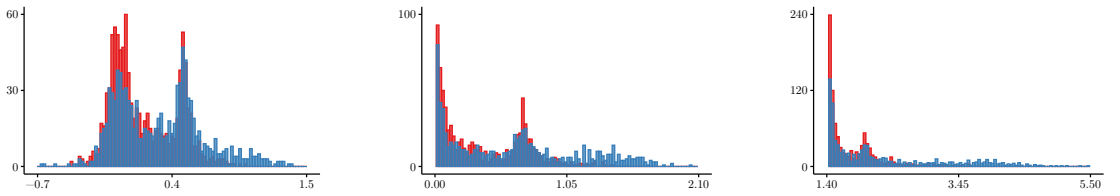
(d) E_S × XF $rs = 0.470, 0.502$)

Figure 9.1: Scatter plots for spatial and decision dynamic measures in Experiment 3. Perceptual competition represented by the circles and solid lines; phonological competition by the crosses and dashed lines. Pearson's r given in each subcaption; perceptual competition given first

Table 9.3 Summary of bimodality statistics per competition type in Experiment 3

Competition	**Measure**	**Bimodality statistics**		**Convergent?**
		b	*HDS* p	
Perceptual	*AUC*	0.556^{*}	$< 0.001^{\dagger}$	✓
	MD	0.663^{*}	$< 0.001^{\dagger}$	✓
	PL	0.643^{*}	$< 0.001^{\dagger}$	✓
Phonological	*AUC*	0.434	$< 0.001^{\dagger}$	✗
	MD	0.622^{*}	$< 0.001^{\dagger}$	✓
	PL	0.728^{*}	$0.010^{\dagger}$	✓

Note. An asterisk (*) or a dagger (†) denote bi- or multi-modality. The 'Convergent?' column denotes whether the tests agree (✓), or disagree (✗).



(a) AUC (b) MD (c) PL

Figure 9.2: Histograms for spatial measures in Experiment 3. Perceptual competition trials represented in red; phonological competition in blue

were not available for testing. There is also support for this idea in the literature – M. W. Smith, Sharit and Czaja (1999) reported that there are clear age effects observable in mouse usage (moving and clicking behaviours, as here).

As the more reliable statistic, HDS was examined more closely across measures. Supported by a visual inspection, this suggested that AUC seemed to be the most bimodal measure. To investigate the cause of bimodality further, a point midway between the two peaks was selected (0.25). Each trial was then categorised as being above or below this point (i.e., belonging to Mode 1 or Mode 2), separately for perceptual and phonological competition trials[3]. Each participant therefore contributed a maximum of 32 trials per condition to each mode. The proportion of trials a participant contributed to each mode was then examined to see if it correlated with age: the hypothesis being that Mode 2 contained largely older participants, and Mode 1 largely younger participants. This was assumed as it was thought that a larger AUC had implied more inefficient responding, more typical of the behaviour of an older person (M. W. Smith et al., 1999). Mode 1 was expected therefore to correlate negatively with age, and Mode 2 positively. However, neither group showed any significant correlation (all rs -0.086 to -0.017). On average, participants contributed 60.4% (SD = 39.1%) of their perceptual trials, and 43.6% (SD = 29.4%) of their phonological trials to Mode 1. To Mode 2, participants gave an average of 34.0% (SD = 36.6%) of their perceptual trials, and 49.6% (SD = 27.8%) of their phonological trials. Mode 1 consisted of 638 perceptual and 460 phonological competition trials; Mode 2 consisted of 359 (perceptual) and 524 (phonological) trials. Mode 1 therefore constituted 55.4% of trials. Given this pattern of data, no valid analysis of the trajectories could be performed (though this was revisited in Experiment 4 – see Section 10.3, p. 133).

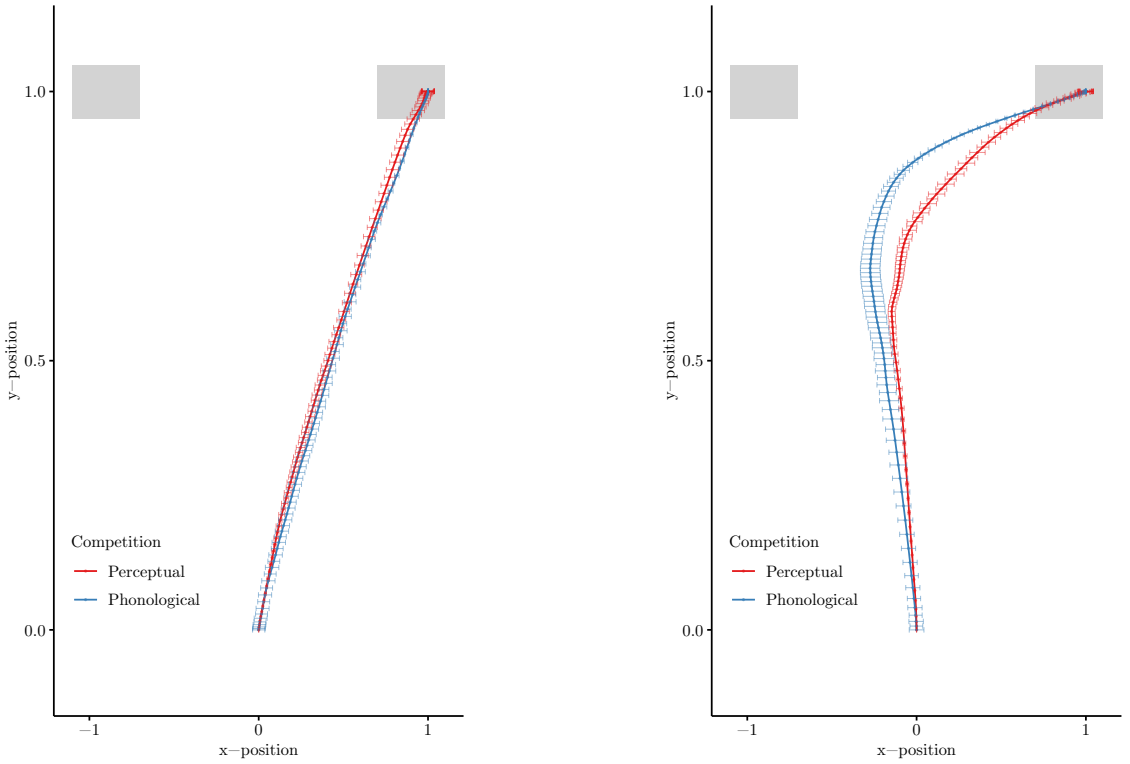
Such a simple explanation of the bimodality did not hold. Instead, with a view to comparing perceptual and phonological trials collapsed over mode, paired-samples t-tests were performed on the data to test if, for any given participant, they were consistent in the number of trials they contributed to either mode. For example, if participants contributed 40% of their perceptual trials to Mode 1, would they contribute the same proportion of phonological trials?

Unfortunately, this was not the case. Participants varied in the proportion of perceptual and phonological competition trials they committed, both in Mode 1 ($t(32) = 7.24$, $p < 0.001$, $d = 0.340$; where the proportion of perceptual trials > phonological trials) and Mode 2 ($t(32) = 6.90$, $p < 0.001$, $d = 0.358$; where the proportion perceptual < phonological).

Averaged participant trajectories

Although the intended trajectory analysis could not be performed, for each mode, the averaged trajectories were imaged. Trajectories consisted of participants' x and y co-ordinates at each time step, aggregated by condition per participant. These were then further collapsed per x- and y-position at each time bin per condition, leaving two averaged trajectories per mode. As such, they illustrated the 'typical'

[3]Each set of competition trials was examined, but by coincidence, a cut off of 0.25 was appropriate for both sets



(a) Mode 1

(b) Mode 2

Figure 9.3: Averaged participant trajectories plotting x- against y-position for Mode 1 and 2 trials in Experiment 3. Each point represents a single time bin. Error bars represent $\pm$ 1SE in the x-position

trial response for the 'typical' participant experiencing that type of competition in that mode. Trajectories can be seen in Fig. 9.3 (p. 122). Whereas Mode 2 trajectories showed the intended competition effect, with movement in the phonological competition condition towards the competitor, Mode 1 showed no such difference. In this condition, participants in both conditions moved towards target directly, seemingly without experiencing competition from the phonological competitor.

9.3.2 Measuring competition with mouse tracking

Although the data in Experiment 3 showed strong bimodality, an important consideration was which measure gave the strongest competition effect. At least on the Mode 2 trials, the trajectory data suggested that the competition was lexical in nature, and so comparison of the two competition types was still valid to perform, setting aside the issue of bimodality. Furthermore, the fact that perceptual trials were more likely to belong to Mode 1, and phonological trials to Mode 2, further suggested there was a 'valid' competition effect to explore. Finally, the effect of collapsing over mode would act only to *weaken* the competition effect, not strengthen it artificially.

The data were therefore once more collapsed over modes, and perceptual and phonological competition trials were compared in a series of paired-samples t-tests

Table 9.4 Summary of phonological/perceptual competition *t*-tests and effect sizes for each measure in Experiment 3

Measures	***t***	***p***	***d***
AUC	6.88	$< 0.001^{***}$	0.586
IT	0.847	0.403, *NS*	0.035
MD	13.8	$< 0.001^{***}$	0.917
PL	17.4	$< 0.001^{***}$	1.43
RT	9.01	$< 0.001^{***}$	0.384
E_S	9.17	$< 0.001^{***}$	0.796
XF	7.68	$< 0.001^{***}$	0.433

Note. $df = 32$. Three asterisks (***) denotes significance below the 0.001 level.

(one for each measure). These are summarised in Table 9.4 (p. 123), along with an appropriate effect size measure to allow comparison across measures (d; Dunlap et al., 1996).

One potential problem of performing these tests was that the data were very clearly skewed, violating the tests' assumption of normality. However, with a sample of more than 30, deviations from normality are not considered to be problematic, and common tests of normality are also overly conservative, i.e., sensitive to even small deviations from normality (Ghasemi & Zahediasl, 2012). Moreover, researchers have reported in simulations of extremely non-normal data that *t*-tests are robust, given samples of such size. Lumley et al. (2002) report that it is a commonly-held, but false, belief that the validity of a *t*-test is dependent on the distribution of the data. The tests were therefore performed without reference to the skewed distribution.

The differences across all measures were significant (all ts ≥ 6.88, ps < 0.001, all ds ≥ 0.389), except for IT ($t(32) = 0.847$, $p = 0.403$, $d = 0.035$). This confirmed that all the measures in Experiment 3 were in some way sensitive to competition, and that 'competition' was not occurring due to selectively later initiation. Of the three spatial measures (AUC, MD, PL), PL showed the strongest effects, and AUC the weakest. On the decision dynamic, E_S was more sensitive than XF, although both were weaker than either MD or PL. RT showed the weakest effect size, as expected (see Section 8.2.2, p. 98; Maldonado et al., 2019).

9.4 Discussion

Experiment 3 had two main objectives. The first was to develop procedures for running mouse tracking analyses and testing the mouse tracking data. The second aim was to find evidence for lexical competition in mouse tracking, and replicate the findings of Spivey et al. (2005). On both of these aims, there was some success. The analyses shown above (Section 9.3, p. 116) clearly demonstrated that running further mouse tracking experiments was a viable proposition.

The second aim was also met, albeit with more limited success. Whilst a competition effect was demonstrated, it was possibly not 'lexical'. Secondly, the trajectory analysis could not be run. Finally, the data were not unimodal, in contrast to Spivey et al.'s work. All of these problems were linked by this failure to replicate Spivey et al.'s unimodal distribution.

The discussion below reflects on the following. First, the design choices made in Experiment 3 will be considered. The second section will cover the problems with bimodality and the nature of the observed competition. The third will contemplate which mouse tracking measures and analyses are optimal for future work. The final section will map out that future work and conclude this chapter.

9.4.1 The suitability of the design choices

Section 9.1.1 (p. 108) made a series of choices not present in the original work of Spivey et al. (2005). A labelling task was included, a limiter on the mouse speed was set, different stimuli were used, and the number of trials were increased. Having demonstrated a competition effect in Experiment 3, is it appropriate to carry these choices over to future work?

The labelling task

Effect sizes were quite large in Experiment 3. Therefore, whilst the labelling task may have helped boost the effects, it is likely not needed in future work. Although it does have precedence in the literature (e.g., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a), it is not universally used (e.g., Spivey et al., 2005; Weighall et al., 2017). It may therefore be regarded as unnecessary, though this would later be examined in Experiment 4 (Chapter 10, p. 129). Furthermore, it is possible that the introduction of the labelling task was the cause of the bimodal data distribution. Perhaps because they had been familiarised with the words and objects, participants were able to manage competition more efficiently, and hence, it was not observed in the Mode 1 trajectories. This is however speculative – and any, or all – of the design choices in Experiment 3 could have been the cause of the bimodal data.

Speed limiter

The limiter on the mouse speed, forcing participants to move the mouse more slowly, seemed to work as intended. As the parameter has also been used by other researchers (e.g., Feather et al., 2014), it was carried forward to future work.

Stimuli changes

It is an open question whether British English-speaking participants would still show competition effects on American English items. Likewise, the effect of the more slowly-delivered word labels is difficult to quantify. However, there is no obvious reason to change the stimuli from those used in Experiment 3, as a competition effect was demonstrated.

Trial numbers

Experiment 3 provided no obvious reason why the number of trials used should be changed. However, design is a critical factor in the running of psychological experiments, and as will be seen, later novel word work would adapt to the design of Weighall et al. (2017), in order to maintain comparability of effects. This resulted in fewer trials in all later work (Chapters 10 and 12 to 14, pp. 129, 155, 177 and 191).

9.4.2 Bimodality and the nature of competition

The finding of bimodally-distributed data (see Table 9.3, p. 120) was a major disappoint in Experiment 3, and represented another failure to replicate, as Spivey et al. found no evidence of bimodality. Particularly problematic was that this limited the analyses that could be performed on the data – the planned comparison of standardised time bins would not have been valid to perform with such data. Most of the measures here showed bimodality, on two different statistics, which was further confirmed by visual inspection (Fig. 9.2, p. 120).

The cause of the bimodality is difficult, if not impossible to establish. What is more important is the implication it has for the results in Experiment 3, and the extent to which it qualitatively changed the 'competition effect' found – making the comparison to other work in the literature invalid.

On at least a subset of trials (i.e., those in Mode 2), there does seem to evidence of lexical processing (and competition) in Experiment 3. Phonological trials showed a clearly disturbed response when compared to perceptual trials (Fig. 9.3, p. 122). Moreover, paired-samples t-tests suggested that it was more likely that a perceptual trial would be in Mode 1, and a phonological trial would be more likely to occur in Mode 2. It is therefore beyond doubt that there was a 'competition effect' in Experiment 3. All the measures intended to show this effect did so (Table 9.4, p. 123).

The goal of Experiment 3 was to establish procedures for conducting further mouse tracking experiments. Therefore, the *nature* of competition in Experiment 3 does, to some extent, not matter at all. Mouse tracking has been used in many domains, and is sensitive to many effects. The consideration of the 'type' of competition is only important if one considers that the results would be different when detecting a more definitively lexical effect. That is to say, the overall *pattern* of effect sizes and significance levels would not hold under all circumstances. However, there is no reason to expect this. Regardless of the type of effect, it is sufficient that Experiment 3 demonstrated *some* form of competition. On that basis alone, it is possible to contemplate what are the optimal measures and analyses for future studies.

9.4.3 Optimal mouse tracking measures and analyses

Establishing a smaller set of measures was an important aim of Experiment 3. Analyses showed that, although all the mouse tracking measures (excluding IT) were sensitive to competition, the degree of correlation with a dynamic was high enough that each measure could not be said to be truly different (Fig. 9.1, p. 119). There was

also a range of effect sizes (ds: 0.384–1.43) – despite the high degree of correlation, clearly there were differences in how sensitive particular measures were. Individual measures were therefore considered to be more or less suitable for carrying over to future experiments. The three 'groups' of measures are each discussed below, in addition to the distributional and trajectory analyses.

Distributional analyses

The bimodal response profile of Experiment 3 emphasised the need to conduct these analyses. Furthermore, it was already established as best practised in the mouse tracking literature (e.g., Freeman & Dale, 2013; Spivey et al., 2005). Of the two measures, Experiment 3 confirmed the argument put forward by Freeman and Dale (2013): b appears to be insensitive under some circumstances (e.g., in Experiment 3, not detecting the bimodality of the phonological competition trials on the AUC measure – confirmed by HDS and visual inspection). It was considered that the best approach was to report both statistics, however, until a unimodal pattern of data was observed.

The decision dynamic

The certainty of a participant's decision was measured with two indices: x-position flips (XF), measuring the inconsistency in horizontal movement, and sample entropy (E_S), measuring stability of trajectory over windows of time. Of the two, XF was the cruder, but more intuitive measure; E_S provided a more sophisticated, and more sensitive measure, but relied on complicated underlying mathematics (Richman & Moorman, 2000; Calcagnì et al., 2017).

Given its increased sensitivity, E_S should clearly be preferred moving forward ($d_{E_S} = 0.796$; $d_{XF} = 0.433$). However, a more fundamental question should be asked of the decision dynamic: is it required at all? The answer is probably not: both effects were smaller than two of the three spatial measure effects. Additionally, what the measures themselves index – instability or uncertainty in the decision making – was easier to understand when shown visually, such as by the trajectory analysis. For these reasons, no further experiments reported measures for this dynamic.

The spatial dynamic

As in other work, the spatial dynamic here produced the largest effect sizes (e.g., Maldonado et al., 2019). Path length (PL) produced by far and away the strongest effect size ($d_{PL} = 1.43$) – much stronger than the other two measures (area under curve (AUC), $d_{AUC} = 0.586$; and maximum deviation (MD), $d_{MD} = 0.917$). For this reason alone, the measure could be chosen. However, there is another factor favouring PL.

Two of these spatial measures measure the difference between the actual mouse path, and an idealised, optimal trajectory. AUC measures the area bound by these two lines (Fig. 8.3a, p. 98), MD, the point at which the trajectory is furthest from it (Fig. 8.3b, p. 98). PL, by contrast, measures the length of the trajectory itself: it does not rely on a contrast to a hypothesised 'optimal' trajectory. This is a

strength of the measure if one considers the 'type' of responses that were observed in Experiment 3.

Consider two plausible response strategies. One participant slowly moves the mouse up the screen, waiting for enough information to respond – which will come either early (in the case of perceptual competition) or late (in the case of phonological competition). All three measures will show differences across these conditions, as the vertical part of the response is larger on phonological competition trials, thus increasing the AUC, MD and PL.

A second strategy is more problematic. This participant, knowing that the response options are in predictable locations, moves the mouse up to the top of the screen, aiming to reach, as quickly as possible, a point equidistant from the two images. S/he can then move the mouse quickly either left or right to respond to an appropriate target. This creates a trajectory shape something like an inverted 'L'. As s/he is pre-disposed to anticipatory responding, occasionally, s/he opts for the wrong target – which s/he then has to correct, creating something more like a trajectory shaped more like a 'T'. With the increased cognitive load on phonological trials, this participant makes more of the 'T'-type responses in this condition.

Comparing such 'L'-type with 'T'-type responses, AUC and MD would show no difference, because in both cases they measure the distance between the vertical portion and an idealised straight-line response from origin to target. By contrast, PL can easily tell these trials apart – going over one 'arm' of the 'T' obviously increases PL. Although the above pattern of responding is hypothetical, and not necessarily observed in real data, it is presumably due to factors like this that PL shows increased sensitivity when compared to AUC and MD.

The temporal dynamic

Whilst it produced a weak effect size, RT is a measure that is ubiquitous in psychological research. Moreover, it acted as a 'check' on the spatial measures, and made sense to report, given that data from this measure was in any case analysed as part of the trajectory analysis. Likewise, IT was another useful 'check' measure, as it eliminated the possibility that differences in RT or the spatial measures were driven by differences in IT. Therefore, it was decided that both measures would be reported in future work.

Trajectory analyses

Although the standardised time bin analysis could not be performed, the usefulness of looking at the trajectory data was obvious in Experiment 3. Mode 1 and Mode 2 were most clearly distinguished by examining the quite different trajectories they gave rise to. Also, the averaged participant trajectories provided a useful visualisation of the spatial differences reported in the data. This, combined with the fact that trajectory analyses are the most frequently reported measure in the mouse tracking literature, made it necessary to include them in future work. In later experiments, the analysis of time slices was also anticipated to suggest differences between novel and familiar words, and the competition they produced.

9.4.4 Conclusions and future work

In conclusion, Experiment 3 established that IT, PL and RT should the favoured measures for future work. For the distributional analysis, whilst HDS was favoured, until a unimodal distribution had been observed, it was hard to draw firm conclusions (although b showed clear deficiencies). It was decided that both measures were to be reported, and one eliminated later, if possible. Lastly, future work would also plot the trajectories: either of each mode (in the case of bimodal data), or against standardised time bins, as appropriate.

However, the design in Experiment 3 was not the same as the design that has been used in novel word learning research (e.g., Weighall et al., 2017). The stimuli were also different. Experiment 4 would therefore use an altered design and stimuli set. This is presented in Chapter 10.

CHAPTER

TEN

EXPERIMENT 4
THE ROBUSTNESS OF MOUSE TRACKING EFFECTS

10.1 Introduction and rationale

In Experiment 3 (Chapter 9, p. 107), mouse tracking successfully demonstrated a competition effect, partially replicating Spivey et al. (2005). Additionally, processes and procedures were developed for designing, running and analysing mouse tracking experiments. However, the design and stimuli used in Experiment 3 (and Spivey et al., 2005) was not well reflected in the novel word literature (e.g., Weighall et al., 2017). Additionally, the conclusions reached in Chapter 9 were made on the basis of a single experiment, with an unusual sample, and an atypical dataset (insofar as it was bimodal). Whilst the conclusions were valid in the context of that experiment, they were not necessarily solid with respect to future work. Experiment 4 was therefore an opportunity to solidify the conclusions (i.e., regarding the optimal measures), and test the robustness of mouse tracking with a new design and stimuli set. This could then be fed forward to novel word learning experiments (Chapters 12 to 14, pp. 155, 177 and 191).

10.1.1 Changes in Experiment 4

Comparing Weighall et al. (2017) and Experiment 3, the most significant difference in design was that items were not repeated in the novel word learning paper. The implication of this is that the labelling task, where participants learnt that the picture of the DOLLAR was to activate the phonological representation 'dollar', not 'money', 'banknote', bill', etc., could not be used. Additionally, a single presentation would reduce the number of trials, and therefore also, the statistical power. This meant that choosing the correct measure (i.e., that with strongest effect size) was doubly important, another reason to test the conclusions reached in Experiment 3 further (Section 9.4.3, p. 125).

The other consequence of Weighall et al.'s single presentation design choice was that each trial stood alone. Whilst the items used by Spivey et al. (2005), and in Experiment 3, were photo-realistic, Weighall and colleagues used cartoons, of lower resolution (200 × 200 pixels, though expanded in Experiment 4 to 300 × 300 for

consistency with Experiment 3). It may have been the case that these smaller, 'less realistic' cartoons activated their lexical labels less strongly, and so produced less competition – an undesirable effect which could then be compounded by a single presentation design. Additionally, they were not always concrete objects as in Experiment 3 (cf., 'robber', 'kitchen', Fig. C.1, p. 217; Table C.1, p. 218). Another difference was that rather than simply saying the words (preceded by 500ms of silence), the speaker in Weighall et al.'s stimuli spoke "Click on the X" (where 'X' was the target's label). These were delivered naturalistically, instead of with a token carrier phrase of a specific length then spliced onto the word. Consequently, length of the carrier phrase was somewhat variable (see Fig. C.2, p. 217, although not all these items were used in Experiment 4). The combination of no labelling task and the single presentation meant that each object had to effectively activate its representation, and it was not immediately obvious that Weighall et al.'s stimuli[1] would in a mouse tracking experiment. The stimuli themselves were therefore also of interest, in addition to the *Competition* variable.

In summary, Experiment 4 therefore had five aims:

1. To demonstrate that the mouse tracking competition effects in Experiment 3 were robust to design changes;

2. To further test the specific conclusions from Experiment 3, with respect to the optimal mouse tracking measures;

3. To look again at bimodality, and consider whether it was a feature of the design, or an artefact encountered only in Experiment 3;

4. In the event of finding a unimodally-distributed data set, to study and perform the standardised time bin analyses which could not be performed in Experiment 3;

5. To establish the influence of the stimuli on any observed competition effects. Was it the case that cartoon stimuli and variably spoken stimuli would significantly attenuate a lexical competition effect?

To this end, analyses in Experiment 4 were conducted as follows. Firstly, with respect to the measures, area under curve (AUC), initiation time (IT), maximum deviation (MD), path length (PL), response time (RT) were analysed. Attention was again paid to the pattern of effect sizes: Experiment 3 suggested that PL was stronger than AUC and MD, and that RT would be a weaker measure than any of the spatial measures. IT was anticipated not to show a difference. Distributional analyses were to be performed with both the bimodality coefficient, b, and Hartigans' dip statistic (HDS), although with a view to cutting b, as it had previously shown itself to be insensitive to demonstrably bimodal data (cf., Table 9.3 and Fig. 9.2, p. 120). Finally, trajectory analyses would illustrate the response profile. In the event of finding unimodal data, the trajectories (collapsed by trial and participant) for each condition would first be imaged – confirming that the measures indicated the

[1]Kind regards to Anna Weighall for sharing her stimuli set.

intended effect (i.e., of displacement towards a competitor). A standardised time bin analysis would then be carried out, comparing the x-position at each time slice across conditions. Positional differences in the x direction would indicate differing levels of attraction to a competitor. Analysis of the y-position data were also considered, but not performed, due to difficulty in matching those data to a relevant theoretical construct.

In the event of bimodal data, the trajectories were to be processed as in Experiment 3, with separate imaging for each mode, but otherwise, no analyses performed.

Driving all these analyses were the two independent variables of interest: *Competition* (*Perceptual, Phonological*), as in Experiment 3, and a new variable, *Stimuli* (*Cartoon, Photo*). *Cartoon* stimuli used pictures and audio from Weighall et al. (2017), and *Photo* used stimuli from Experiment 3.

10.2 Methods

10.2.1 Participants

In Experiment 4, data were collected from 35 undergraduates (five male, $M_{\text{Age}} = 21.3$ years, $SD_{\text{Age}} = 7.82$ years, 29 monolingual, 29 right-handed), all fluent in English. Participants were all tested in a quiet laboratory environment. All participants had not participated in any previous experiments. All were free of any confounding disorders (e.g., sensory, learning or language difficulties), or had corrections to normal (e.g., by wearing eyeglasses). Participants all ordinarily used the mouse with their right hand.

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, or in exchange for course credits.

10.2.2 Apparatus and Materials

The audio and visual stimuli for the *Photo* block were re-used from Experiment 3, with the addition of four further trials to bring the total up to ten in each of the perceptual and phonological competition conditions. These stimuli were photo-realistic depictions of common-place objects, and were all concrete nouns. The *Cartoon* block, used a mix of less-concrete (e.g., KITCHEN) and concrete (e.g., ONION) nouns, which were depicted as cartoon-like clip art illustrations. However, Weighall et al. (2017) reported that all of these illustrations had $\geq 80\%$ naming agreement amongst 10 adults. Furthermore, *Cartoon* items were matched on written and verbal frequency, concreteness, familiarity and imageability, according to the MRC Psycholinguistic Database (Wilson, 1988). A full list of the stimuli used can be seen in Table C.1 (p. 218).

The audio files used in the *Cartoon* block were also slightly different, as again, they had been taken from Weighall et al. (2017), rather than being created locally. Instead of consisting of a single word, preceded by 500ms of silence, the audio files were all the phrase "Click on the X" (where 'X' was the target's label). This meant that the object labels were heard slightly later in the *Cartoon* block, as the average

length of this carrier phrase was 631ms (SD = 64ms). The carrier phrase was of variable length due to the recordings having been made by naturalistic speaking, instead of splicing a single phrase onto the recorded object labels. The stimuli were of slightly poorer quality, with a small amount of audible room echo – by contrast, *Photo* stimuli had been recorded in a sound attenuating booth. The final difference was that the speaker was female, and spoke with a southern English accent, contrasting with the male northern voice used in the *Photo* block, and Experiment 3.

Variables such as testing apparatus (e.g., computer screen, mouse, etc.) and picture size or location did not vary from Experiment 3. Items taken from Weighall et al. (2017) in the *Cartoon* block were therefore re-sized from 200 × 200 pixels to the 300 × 300 pixels used previously. The MouseTracker script parameters (e.g., speed) were similarly maintained across both experiments.

10.2.3 Design

Participants progressed through a total of 40 trials, which were blocked depending on the *Stimuli* variable. Twenty trials were like those used in Experiment 3 and Spivey et al. (2005) – this was the *Photo* block. By contrast, the remaining 20 trials were created with stimuli taken from Weighall et al. (2017); these were the *Cartoon* block. Which block a participant saw first was randomised. Within each block, participants saw 10 phonological (e.g., CAMPER and CAMEL, disambiguating after the first syllable), and 10 perceptual (e.g. MITTEN and ANGEL, disambiguating at the first phoneme) competition condition pairs interleaved. This created a 2 × 2 within-subjects design: *Stimuli* (*Cartoon, Photo*) × *Competition* (*Perceptual, Phonological*).

Each pair of items appeared once, and only in its pairing (see Table C.1, p. 218). This was a design choice which reflected the procedure used by Weighall et al. (2017). Left and right target presentations were counterbalanced across trials (5 trials per design cell) instead of repeating items, as had been done in Experiment 3.

10.2.4 Procedure

To ensure a valid comparison, as little as possible was changed in the procedure between Experiments 3 and 4. Participants were again instructed to click on a picture for a word that they heard as quickly and accurately as possible. Unlike in Experiment 3, in Experiment 4, participants did not first proceed through a labelling task – instead, they went straight into the experimental trials. They would began with either the *Cartoon*, or *Photo* block first, and then ended with the other block of items. As there were only a small amount of trials, participants finished the experiment very quickly (in around 5–10 minutes). Participants were debriefed and thanked for their time at the end of the experiment.

With respect to the measures, reported here are AUC, IT, MD, PL, and RT, in addition to the distributional and trajectory analyses. The distributional analyses still make use of both the bimodality coefficient (b; SAS Institute Inc., 2018), and Hartigans' dip statistic (HDS; Hartigan & Hartigan, 1985). The inclusion of all the spatial measures, and both of the sensitive bimodality statistics, was an opportunity to confirm the findings of Experiment 3.

Processing of data and exclusions

Procedures for processing the data and excluding trials and participants followed that in Experiment 3 (Section 9.2.4, pp. 114 and 115).

Trials were remapped to the right side of the screen, and the data resampled to even steps of 20ms by linear interpolation. Left and right-orientated trials were collapsed. Trajectories were spatially and time-normalised (Dale et al., 2007; Spivey et al., 2005). The period before initiation was removed and the timestamps were reset to zero, meaning that RTs were indicative of movement, not overall response, time.

Each of the 35 participants completed 40 trials, resulting in an initial data set of 1400 trials. Firstly, 22 trials (1.57% of total) were removed due to participant error, leaving 692 *Cartoon* and 686 *Photo-realistic* trials. The dataset was then further divided into perceptual and phonological competition trials, resulting in four subsets, for trimming per experimental condition. Although Freeman and Dale (2013) urged caution about *SD* trims, as it had been performed in Experiment 3, it was again performed here.

Following a trim from the $M \pm 3SD$ on the AUC, MD, PL and RT data, there remained 326 perceptual, and 323 phonological, competition *Cartoon* trials, and 326 perceptual, and 324 phonological, competition *Photo* trials. Therefore, in total, 101 trials (7.21%) were removed, leaving 1299 trials (92.8%).

All four sets of trials were then examined to see if any participants were contributing less than seven trials out of a maximum of ten in any cell of the design[2]. This led to the elimination of one participant, and further 23 trials. Each subject therefore submitted an average of 37.5 trials out of a maximum of 40 (93.8% of total).

10.3 Results

Mouse tracking data were collected in MouseTracker (Freeman & Ambady, 2010). All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016), and the mouse tracking data were processed with `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff et al., 2020).

10.3.1 Descriptive statistics

Table 10.1 (p. 134) gives a summary of the condition descriptive statistics for Experiment 4. Across both blocks of stimuli, there was more temporal and spatial disruption when responding to a phonologically competing pair of objects. However, this did not appear to be due to slower ITs, which were broadly similar across all trials. This was consistent with Experiment 3. Interestingly, *Photo* trials appeared to induce less efficient responding overall, even on perceptual competition trials. This was not expected, although this increase in perceptual competition may

[2]It was impossible to use Experiment 3's cut-off of 75% as each participant contributed 10 trials per condition. To keep as many trials (and participants) as possible, the cut was rounded down rather than up.

Table 10.1 Summary of descriptive statistics per level of *Competition* and *Stimuli* for each measure in Experiment 4

Stimuli	Measures	Competition			
		Perceptual		*Phonological*	
		M	*SD*	*M*	*SD*
Cartoon	*AUC*	0.179	0.174	0.280	0.152
	IT (ms)	380	305	381	327
	MD	0.307	0.212	0.482	0.244
	PL	1.72	0.201	2.04	0.339
	RT (ms)	1349	300	1462	289
Photo	*AUC*	0.207	0.168	0.346	0.137
	IT (ms)	395	318	378	320
	MD	0.334	0.205	0.574	0.234
	PL	1.76	0.219	2.19	0.403
	RT (ms)	1406	278	1581	268

Note. The units for area under curve, maximum deviation and mouse path length are arbitrary.

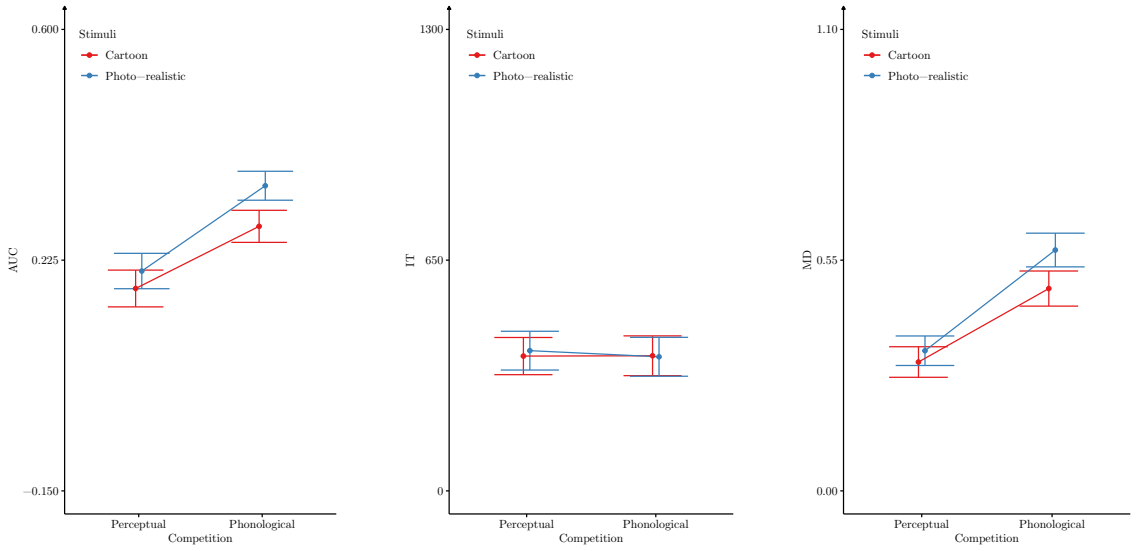
have been due to stronger activation of the competitor, regardless of condition. This difference was particularly surprising if one considered that the name of the target was also heard later on *Cartoon* trials, due to the length of the carrier phrase. This meant that participants responding to *Photo* stimuli could have responded earlier – but they did not appear to. Means plots for each measure are shown in Fig. 10.1 (p. 135).

Correlations between mouse tracking measures

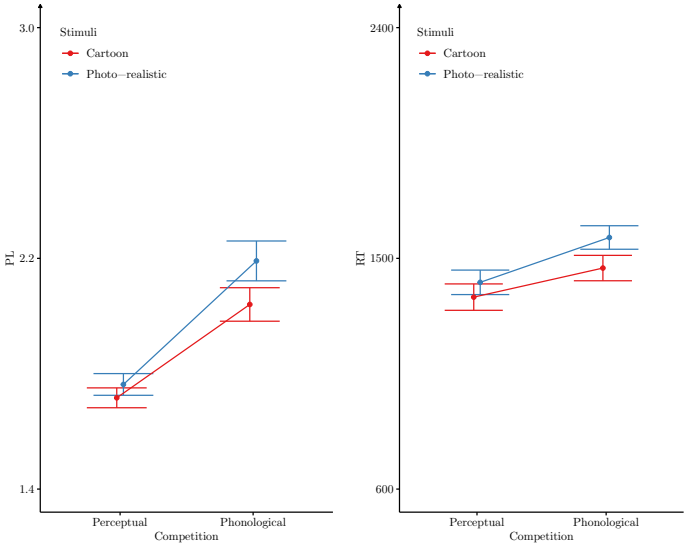
Correlations for Experiment 4 are shown in Fig. 10.2 (p. 136). As in Experiment 3, the spatial measures again exhibited very strong positive correlation, with AUC × MD showing the strongest correlation, and AUC × PL showing the weakest (cf., Fig. 9.1, p. 119). This was consistent across both experimental blocks and trial types.

Distributional analysis

Bimodality statistics were again calculated for all three spatial measures. Unlike in Experiment 3, b and HDS did not show convergence, except for the AUC measure (see Table 10.2, p. 137). Only AUC phonological competition *Cartoon* trials showed evidence of bimodality on both b and HDS (all other HDS $ps \geq 0.615$, *NS*). b was more variable than HDS and suggested bimodality on several measures. However,

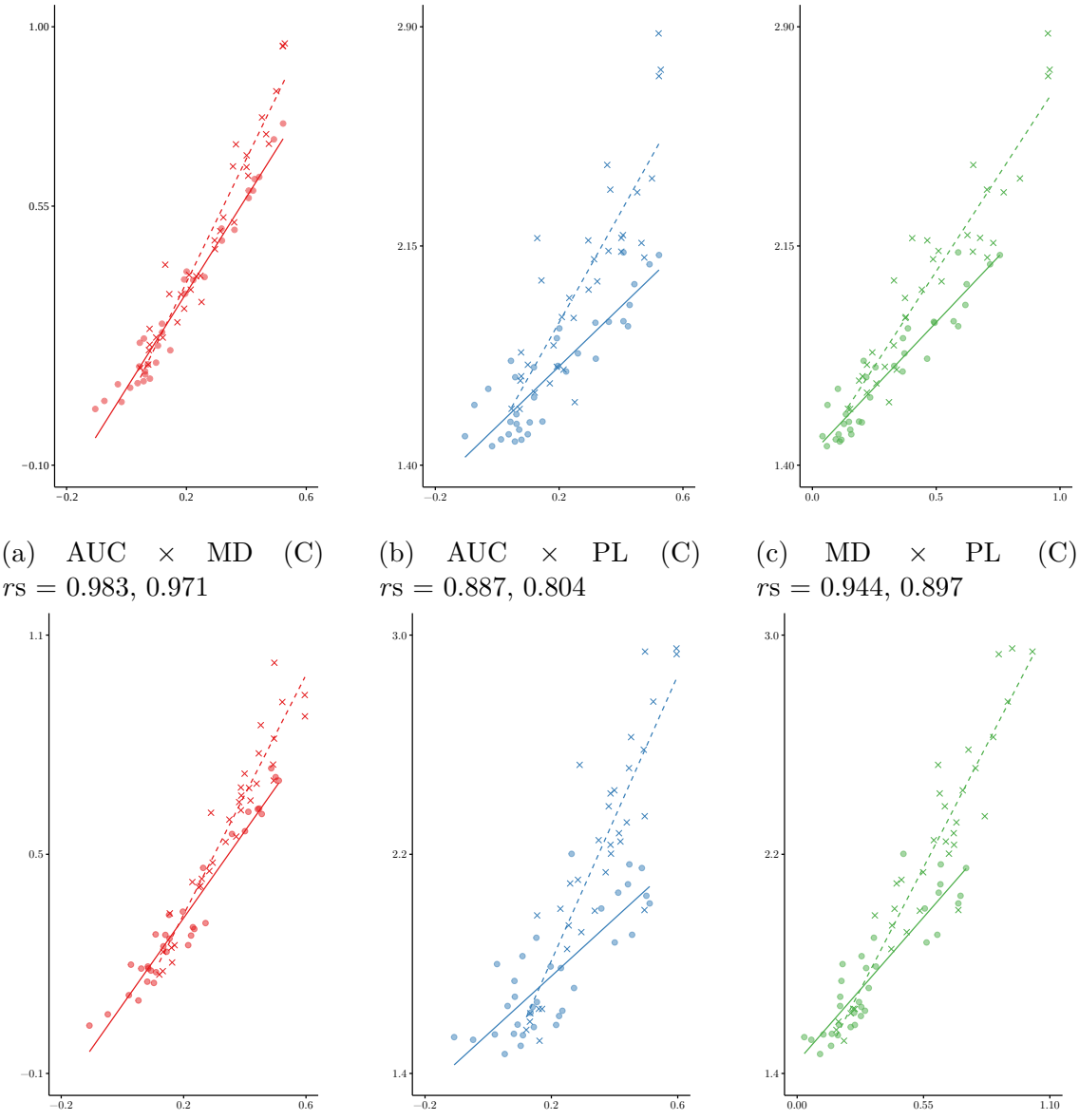


(a) AUC (b) IT (c) MD



(d) PL (e) RT

Figure 10.1: Means plots for each measure in Experiment 4



(a) AUC × MD (C) rs = 0.983, 0.971

(b) AUC × PL (C) rs = 0.887, 0.804

(c) MD × PL (C) rs = 0.944, 0.897

(d) AUC × MD (P) rs = 0.966, 0.991

(e) AUC × PL (P) rs = 0.851, 0.887

(f) MD × PL (P) rs = 0.940, 0.943

Figure 10.2: Scatter plots between spatial measures for each stimuli type. The top row is for cartoon stimuli (C); the bottom photo-realistic (P). Circles and solid lines display perceptual competition trials; crosses and dashed lines phonological competition trials. Pearson's r given in each subcaption; perceptual competition given first

Table 10.2 Summary of bimodality statistics per level of *Competition* and *Stimuli* for each measure in Experiment 4. See also Fig. 10.3, p. 138

Stimuli	Measures	Competition			
		Perceptual		*Phonological*	
		Bimodality statistics			
		b	*HDS* p	*b*	*HDS* p
Cartoon	*AUC*	0.442	0.834, *NS*	0.428	0.010†
	MD	0.600*	0.615, *NS*	0.629*	0.884, *NS*
	PL	0.611*	0.992, *NS*	0.763*	0.540, *NS*
Photo	*AUC*	0.459	0.826, *NS*	0.447	0.087, *NS*
	MD	0.573*	0.992, *NS*	0.620*	0.606, *NS*
	PL	0.688*	0.989, *NS*	0.732*	0.641, *NS*

Note. An asterisk (*) or a dagger (†) denote bi- or multi-modality.

visual inspection of the histograms confirmed what had previously been noted by Freeman and Dale (2013): *b* was biased towards bimodality by skewed data (see Fig. 10.3, p. 138). Both MD (Figs. 10.3b and 10.3e) and PL (Figs. 10.3c and 10.3f) were heavily skewed. This suggested that in future work, HDS should be favoured over *b* when conducting distributional analyses.

Similarly to Experiment 3 however, the histograms (Fig. 10.3 p. 138) showed evidence of skewed data distributions. However, as described in Section 9.3 (p. 123), this is not a problem in samples of this size, or on a *t*-test (Lumley et al., 2002; Ghasemi & Zahediasl, 2012). The literature has also shown that ANOVAs are likewise robust to deviations from normality in the distribution of the data (Blanca et al., 2017). The normality of the distributions were therefore not considered further.

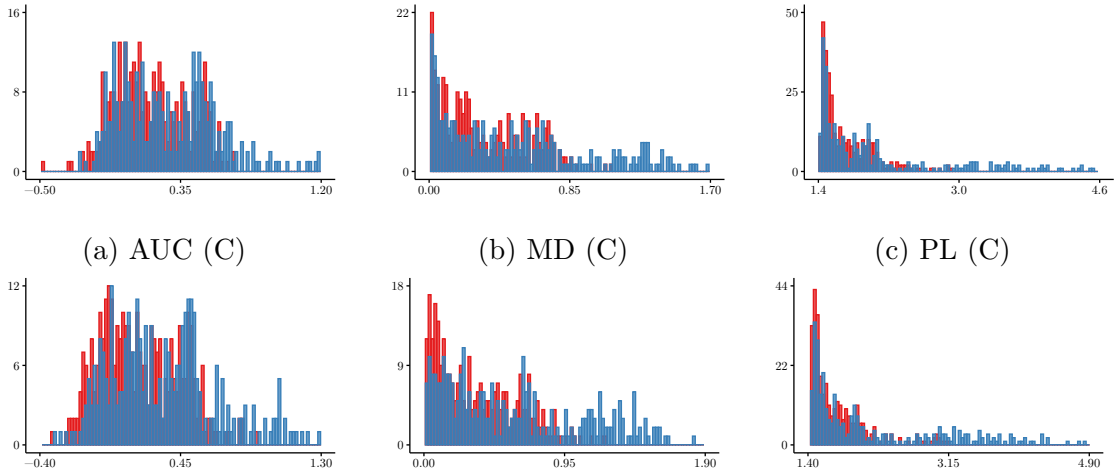
Averaged trial trajectories.

Fig. 10.4 (p. 138) shows the mean participant perceptual and phonological competition trajectories when participants were responding to each stimuli set. All Experiment 4 trajectories were qualitatively similar to those seen in the second mode of Experiment 3, suggesting that the impact of design and stimuli changes between experiments on responding were minor (cf., Fig. 9.3b, p. 122).

10.3.2 Inferential statistics

Mouse tracking measures

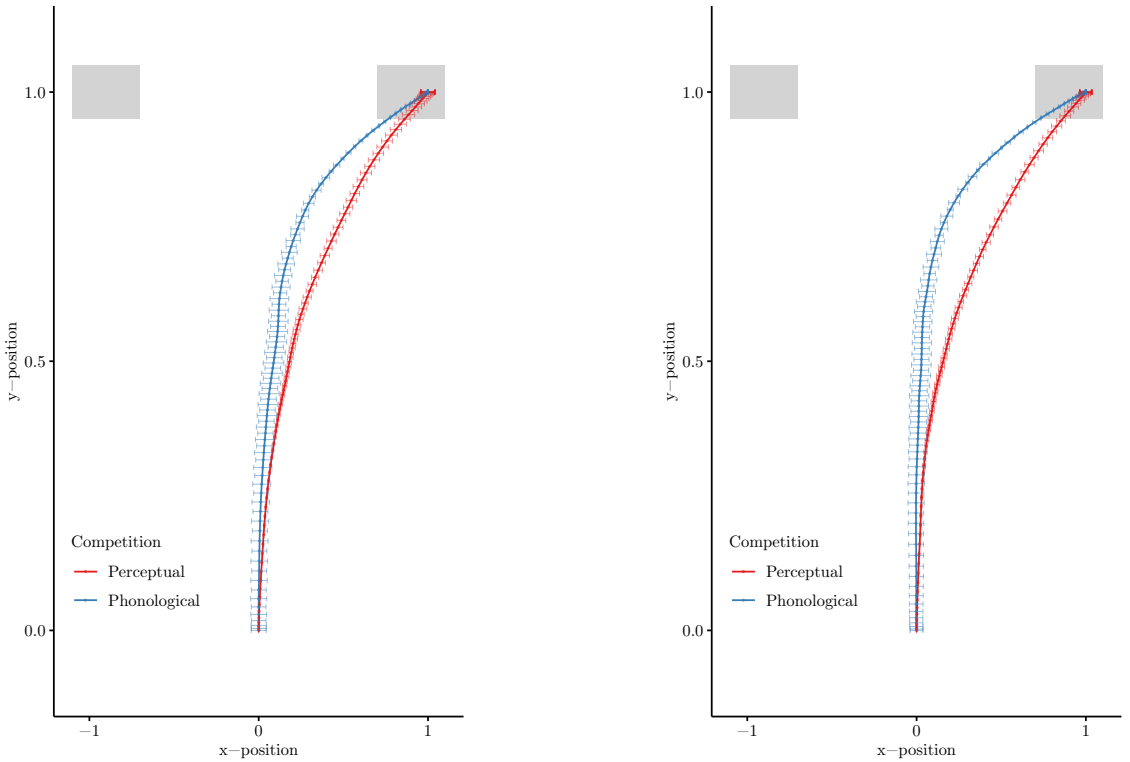
The first block of inferential tests performed were five 2 × 2 repeated-measures ANOVAs (one for each measure, see Table 10.3, p. 140), entering the two independent variables *Competition* (*Perceptual*, *Phonological*) and *Stimuli* (*Cartoon*, *Photo*).



(a) AUC (C) (b) MD (C) (c) PL (C)

(d) AUC (P) (e) MD (P) (f) PL (P)

Figure 10.3: Histograms between spatial measures for each stimuli type. The top row is for cartoon stimuli (C); the bottom photo (P). Perceptual competition trials shown in red; phonological competition trials shown in blue. See also Table 10.2, p. 137



(a) Cartoon (b) Photo

Figure 10.4: Averaged participant trajectories plotting x- against y-position for each stimuli type in Experiment 4. Each point represents a single time bin. Error bars represent $\pm$ 1SE in the x-position

Except for IT, all measures showed main effects for *Competition* and *Stimuli*. Critically, however, no interaction was statistically significant, although MD, PL and RT all showed $p < 0.10$. The lack of a significant difference in any IT comparison provided confidence that the differences found were not due to later initiation.

Before further post-hoc tests were conducted to explore these effects further, the spatial measures' effect sizes were examined. It was noted that PL gave the strongest effects, with MD again giving stronger effects than AUC on the variable of interest for future work, *Competition*. PL gave a *Competition* effect size approximately 160% that of MD. RT produced much weaker effects than any of the spatial measures, as expected (Maldonado et al., 2019).

Although no interaction effect was observed, as the effect was only marginally non-significant, a two blocks of t-tests were performed. The first contrasted perceptual and phonological competition, separately for each stimuli type. The second set compared each stimuli type's perceptual and phonological competition trials.

Comparison of *Phonological* and *Perceptual* competition. The first block of t-tests looked for a competition effect separately for each stimuli type. These tests are summarised in Table 10.4 (p. 141). All measures were sensitive to competition effects (all ts ≥ 4.93, all ps < 0.001). Of all the measures, PL was again the most sensitive, and all the spatial measures were more sensitive than RT.

Comparing the effect sizes for each stimuli type, *Photo* stimuli produced stronger effects (approximately 16% larger on PL, for example). However, the difference between effect sizes for each stimuli type was smaller than the difference in effect sizes across experiments (32% drop in size of d from Experiment 3 to Experiment 4 on *Photo* trials; cf., Table 9.4, p. 123). This suggested that the stimuli differences in Experiment 4 were not as important as the other design changes. Furthermore, it was still the case that PL gave rise to a strong effect ($d = 0.988$) with *Cartoon* trials.

Comparison of *Cartoon* and *Photo* stimuli. Comparisons between *Cartoon* and *Photo* stimuli for each type of competition are shown in Table 10.5 (p. 141). This shows that participants' attention to the competitor was captured more on phonological competition trials. No difference was observed for perceptual competition trials, although in the case of the AUC and PL measures, this was only due to the application of the Bonferroni correction. All effect sizes were quite modest. This confirmed what had been suggested by the other analyses given above: *Stimuli* was not a particularly important variable in designing mouse tracking experiments.

Averaged participant horizontal movement

A unimodal distribution of data allowed for analysis of the x-position against standardised time bins in Experiment 4. The analysis was conducted separately by *Competition* and *Stimuli*. The relevant graphics are shown in Fig. 10.5 (p. 142).

To control for multiple comparisons, a run of time bins was considered significant if they occurred in straight runs of more than eight bins (Dale et al., 2007).

Table 10.3 Summary of *Competition* × *Stimuli* ANOVAs for each measure in Experiment 4

Measure	**Effect**	F	p	η_g^2
AUC	*Competition*	48.6	$< 0.001^{***}$	0.128
	Stimuli	13.5	0.001^{***}	0.022
	Competition × *Stimuli*	2.76	0.106, *NS*	0.004
IT	*Competition*	0.239	0.628, *NS*	< 0.001
	Stimuli	0.040	0.421, *NS*	< 0.001
	Competition × *Stimuli*	0.396	0.534, *NS*	< 0.001
MD	*Competition*	104	$< 0.001^{***}$	0.181
	Stimuli	11.6	0.002^{**}	0.018
	Competition × *Stimuli*	4.12	0.051, *NS*	0.005
PL	*Competition*	107	$< 0.001^{***}$	0.285
	Stimuli	10.8	0.002^{**}	0.027
	Competition × *Stimuli*	3.35	0.076, *NS*	0.008
RT	*Competition*	50.9	$< 0.001^{***}$	0.062
	Stimuli	7.59	0.009^{**}	0.024
	Competition × *Stimuli*	3.56	0.068, *NS*	0.003

Note. $df = (1, 33)$. Three asterisks (***) denotes significance at the 0.001 level, two asterisks (**) at the 0.01 level.

Table 10.4 Summary of phonological − perceptual competition trial *t*-tests and effect sizes by *Stimuli* for spatial and RT measures in Experiment 4

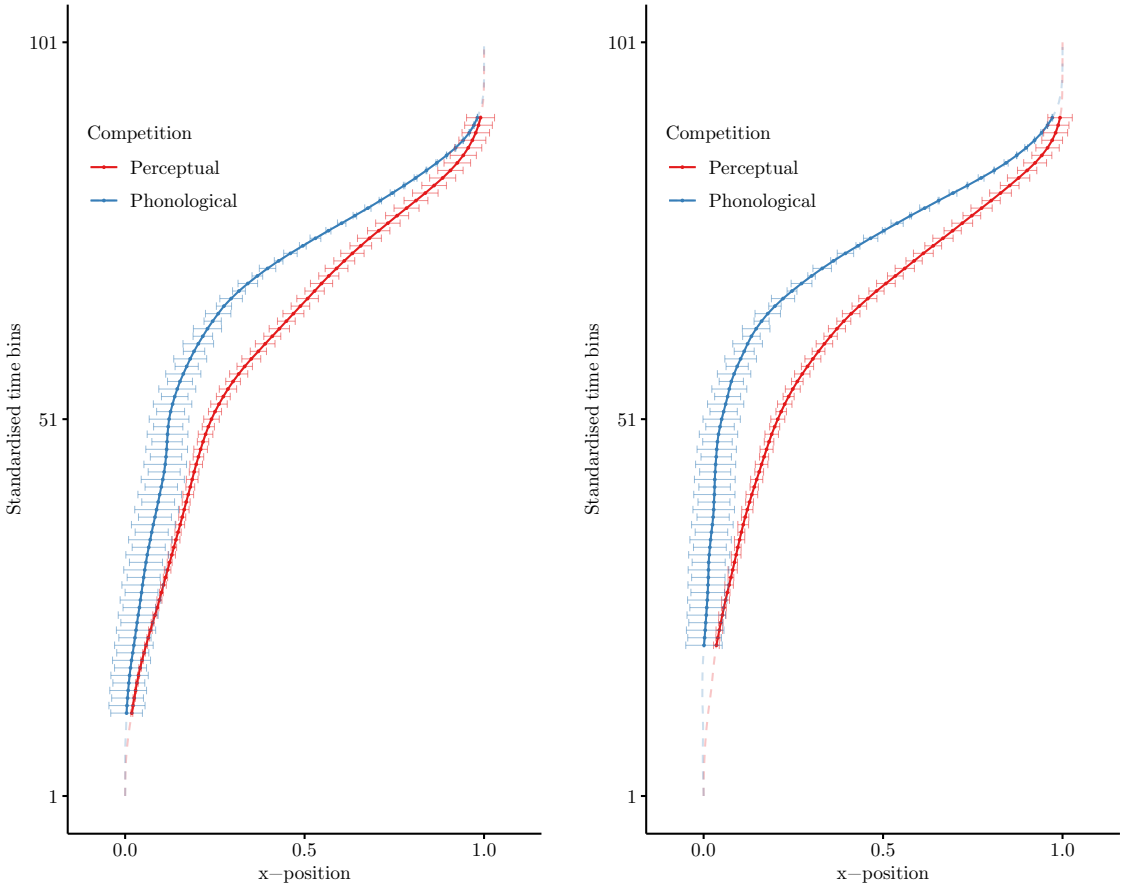
Stimuli	**Measures**	***t***	***p***	***d***
Cartoon	*AUC*	5.68	$< 0.001^*$	0.606
	MD	7.51	$< 0.001^*$	0.748
	PL	8.28	$< 0.001^*$	0.988
	RT	4.93	$< 0.001^*$	0.384
Photo	*AUC*	6.02	$< 0.001^*$	0.890
	MD	8.51	$< 0.001^*$	1.08
	PL	8.15	$< 0.001^*$	1.15
	RT	6.10	$< 0.001^*$	0.640

Note. $df = 33$. An asterisk (*) denotes significance at or below the Bonferroni-corrected *p*-value of 0.013.

Table 10.5 Summary of Photo − Cartoon stimuli trial *t*-tests and effect sizes by *Competition* for spatial and RT measures in Experiment 4

Competition	**Measures**	***t***	***p***	***d***
Perceptual	*AUC*	2.54	0.016, *NS*	0.165
	MD	1.86	0.072, *NS*	0.129
	PL	2.14	0.040, *NS*	0.218
	RT	1.63	0.114, *NS*	0.198
Phonological	*AUC*	3.07	0.004^*	0.454
	MD	3.04	0.005^*	0.384
	PL	2.77	0.009^*	0.400
	RT	3.26	0.003^*	0.427

Note. $df = 33$. An asterisk (*) denotes significance at or below the Bonferroni-corrected *p*-value of 0.013.



(a) Cartoon

(b) Photo

Figure 10.5: Averaged x-position against standardised time in Experiment 4, for each stimuli type. The solid saturated line indicates significantly different positions across conditions in each time bin (paired-samples t-test; $p < 0.05$). Each point represents a single time bin. Error bars represent $\pm$ 1SE in the x-position

Significant differences in the x-position at particular time points was indicative of the differential attraction of the competitor object across levels of *Competition*.

Fig. 10.5 shows, for both *Cartoon* and *Photo* trials, that whereas on perceptual competition trials participants were able to move to their target with only minimal attraction to the competitor object, this was not the case on phonological competition trials. This is indicative of lexical competition. For *Cartoon* stimuli, there were significant differences in the x-position from the 12$^{\text{th}}$ to the 91$^{\text{st}}$ time bin. By contrast, positional differences emerged slightly later in the x-position when participants were responding to *Photo* stimuli. The first significant t-test occurred at the 21$^{\text{st}}$ time bin; competition then persisted again until the 91$^{\text{st}}$ bin.

10.4 Discussion

Experiment 4 had five aims. In summary, these were concerned with replicating and expanding on Experiment 3, by showing that the conclusions reached in that

experiment were robust, and that mouse tracking was robust to changes in design and stimuli. Experiment 4 showed emphatically that the effects and conclusions were robust. Another aim was to look again at the distribution of the data. It was hoped that Experiment 4, unlike Experiment 3, would replicate Spivey et al. (2005) and find unimodal data. This would allow for trajectory analyses to be conducted.

Whatever problem had caused bimodality in Experiment 3, it did not re-occur in Experiment 4, and trajectory analyses were performed. This laid a solid foundation for conducting novel word experiments in Chapters 12 to 14 (pp. 155, 177 and 191).

The detection of unimodality confirmed some ideas suggested by Experiment 3. Firstly, it meant that the bimodality coefficient, which showed itself in both Experiments 3 and 4 to be unreliable and/or insensitive, could be dropped. Future studies would only report HDS, as this was shown to be the most sensitive and robust measure of bimodality (consistent with Freeman & Dale, 2013). As an additional check on the veracity of HDS, the distribution of the data would also still be inspected visually, however.

As to *why* Experiment 3 showed bimodality and Experiment 4 did not, this is unclear. However, given the importance of unimodal data for trajectory analysis, it was deemed prudent not to alter the design for the novel word learning studies.

Unimodality also implied that the effects observed in Experiment 4 were *lexical*, insofar as they are predicted by speech perception models (e.g., Gaskell & Marslen-Wilson, 1997). The Mode 1 trials of Experiment 3 suggested that there was no competition on a substantial number of the phonological competition trials. When, in Mode 2, those trials did show deflection towards the competitor, it was therefore hard to qualify the competition as 'lexical'. Put another way, as the opportunity for lexical competition was present across all phonological trials, the fact that such a large proportion of phonological trials (i.e., in Mode 1, over 50%) showed no competition may have implied that when competition was observed (i.e., in Mode 2) it was not driven by difficulties in lexical processing. By contrast, Experiment 4 produced effects much more consistent with the literature.

Experiment 4 also suggested that the stimuli used did not create significantly differently-sized competition effects (as there was no interaction). Moreover, effect size differences between Experiments 3 and 4 were larger than those observed within Experiment 4, suggesting that design was a more important variable than stimuli. This is considered further below.

10.5 General discussion of Experiments 3 and 4

Two pilots experiments were conducted to answer the questions set out in Section 8.4 (p. 105). It is now possible to answer those questions, with the following conclusions:

1. Mouse tracking is a viable proposition for language and novel word research.
    (a) Procedurally, it is feasible. Whilst analysis of the mouse tracking data is complex, use of packages such as `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff et al., 2020) ease the process.
2. Not all measures and measurement dynamics are necessary.

(a) PL was consistently the strongest spatial measure. HDS was more sensitive than *b*. The decision dynamic provided no more information than that detected by the spatial measures. IT confirmed that data competition effects were not due to different initiation strategies. RT produced weaker but significant effects, and was reported to qualify PL. Trajectory imaging and analyses confirmed that differences in measures were due to selectively greater displacement towards the competitor; time slices gave some view of the time course of competition (emerging, as expected, as predicted by speech perception models, relatively early, e.g., Gaskell & Marslen-Wilson, 1997).

3. The observed competition effects are robust to changes in design and stimuli.

(a) Competition was demonstrated with large effect sizes in two consecutive experiments.

Taking the above, the rest of this section considers the similarities and differences between Experiments 3 and 4 and sets out how the findings were then applied to novel word research.

10.5.1 Conclusions for novel word research

The first and most important consideration for future work is the mouse tracking measures. In this regard, the two experiments spoke as one, as they produced a pattern of effect sizes which was unanimous. The measures to carry forward were IT, PL and RT. HDS was selected over *b* as the test of bimodality, accompanied by visual inspection of the histogram of PL. Trajectory imaging and analysis was also selected to continue in future work. These were particularly useful, for example, in distinguishing the reason for a difference between modes in Experiment 3.

The *Cartoon* stimuli donated by Anna Weighall were favoured over the *Photo* set as the control of its psycholinguistic properties was somewhat better (having being normed; see Weighall et al., 2017), and although it produced numerically weaker effects, these were not statistically weaker. In any case, the practicalities of using this data set also outweighed the slight weakening of the effects.

Experiments 3 and 4 used different designs, and this may have had important effects (possibly in the distribution of data, possibly in the strength of the effects). However, in selecting a design for future work, one must consider also the benefit of methodological closeness to the published design used by Weighall et al. (2017). Rather than a simple comparison of the two designs (as each has its possible flaws), the question should instead be framed 'Does the benefit of changing the design (from that used by Weighall and colleagues) outweigh the cost of methodological distance?'.

Simply put, the answer is 'no'. Whereas effect sizes were stronger in Experiment 3 ($d_{PL} = 1.43$, compared to $d_{PL} = 0.988$ for *Cartoon* stimuli in Experiment 4), in neither experiment was the effect size small enough that it would likely not replicate. A further consideration is that sample size is likely to be as important as the strength

of the effect, as both experiments here used fairly limited samples which could be increased to account for the weaker novel word effects (if they were present).

Thus, the novel word study (Experiments 5–7) proceeded using the stimuli and design of Weighall et al. (2017). The next chapter provides an overview of the study, setting out the relationship between these three experiments.

Part IV

The nature of novel word representations

CHAPTER

ELEVEN

MOUSE TRACKING AND NOVEL WORD LEARNING

11.1 Overview of novel word learning experiments

Following the success of Experiments 3 and 4 (Chapters 9 and 10, p. 107 onwards), and the establishment of a suitable mouse tracking methodology, Experiments 5–7 focussed on novel words. All of the experiments use as their foundation the design and logic of Weighall et al. (2017), adapting that research to mouse tracking.

Weighall et al. used the visual world paradigm (VWP; Tanenhaus et al., 1995), an eye tracking task, to study novel word representations. As discussed in Chapters 7 and 8 (pp. 83 and 93), mouse tracking and eye tracking tasks are similar, but different. In this iteration of the VWP, Weighall and colleagues asked their participants to click on objects whose labels they heard, whilst their fixations to on-screen targets and competitors was measured. Twenty four words were trained the day before testing, and a further 24 words were trained on the day of testing. In brief, they found that the lexical engagement of novel words with their phonological competitors (e.g., between 'biscal' and 'biscuit') was present for words learnt immediately before testing. Furthermore, comparing words learnt on the day of testing and the day before testing, they found that the lexical engagement was indistinguishable (contrary to previous research on sleep, and a complementary systems accounts, e.g., Bowers et al., 2005; Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010). However, consolidation effects were still seen in the recall tasks, where words learnt the day before testing were more readily recalled. Polysomnography data showed that the size of the difference in performance was correlated with components of the sleep cycle.

Data also showed that there were different patterns of activation for novel and familiar words (see Fig. 11.1, p. 150). The activation of familiar words was 'cleaner' – a rapid peak of activation, followed by a rapid decline – whereas the activation of novel words was 'noisy' – the peak of their activation was smaller, and declined slower. The authors therefore concluded that novel words were qualitatively different from familiar words, and that this difference was related to an on-going sleep-based consolidation process. Nevertheless, this work provided a sharp contrast with previous complementary learning systems accounts (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010), and added to the body of evidence showing that novel words are

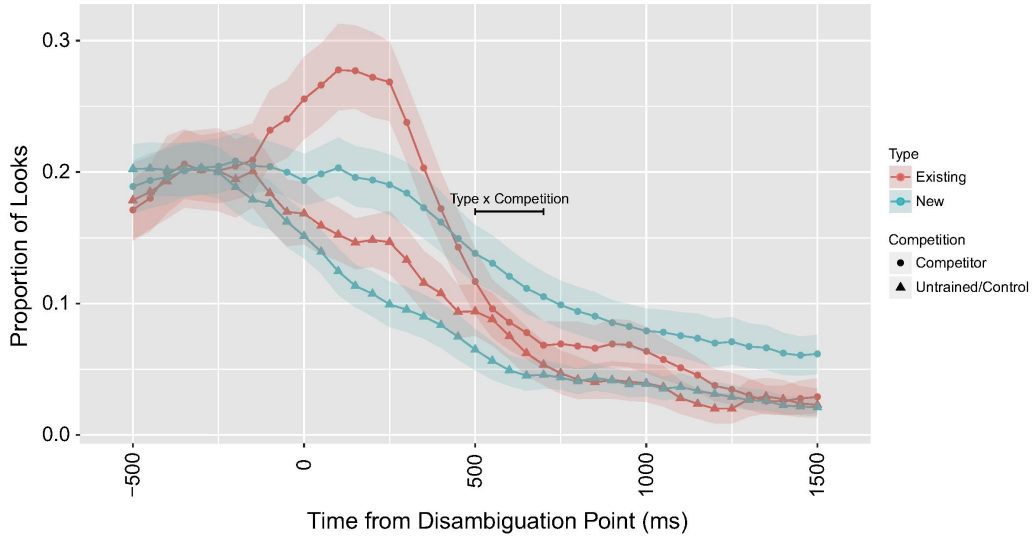


Figure 11.1: Novel ('New') and familiar ('Existing') word competition in Weighall et al. (2017)

capable of immediate (or pre-sleep) lexical engagement (see Chapter 7, p. 83). Although pre-sleep lexical engagement has been demonstrated in several studies, the effect itself is under-explored in the literature. This is addressed in Experiments 5–7.

One of the problems adapting VWP work to mouse tracking is that mouse tracking tests for competition between two objects, whereas in a VWP array, there are typically four objects (though critical comparisons are still made between two objects; e.g., Bartolotti & Marian, 2012; Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Weighall et al., 2017). Weighall et al. used several trial types, outlined below. In all cases, the 'target' was the word that participants heard when they were instructed to click on one of the objects.

- Novel competitor trials, featuring a (i) 'biscuit': a (familiar) target word; (ii) 'biscal': a trained novel competitor (learnt yesterday, or immediately before testing); (iii) 'newspaper': a familiar, perceptual competition distractor; (iv) 'tegwop': an untrained ('super-novel') object and perceptual competitor.
  - For each day of learning, the novel word was either a phonological competitor for the target, or not. Where the novel word was not a phonological competitor for the target, the target had had no competitor trained (e.g., 'guitar', but 'guitas' – its novel competitor – had not been learnt).
  - Fixations to the novel competitor were measured.
- Super-novel target (filler) trials, featuring three familiar objects, and a super-novel target object. These were included to stop participants' attention being biased too much against the novel objects, as participants did otherwise not need to respond to any novel object during the experiment.
  - These trials were discarded as filler trials.

- Familiar competitor trials, featuring a (i) 'candle': a known target; (ii) 'candy' or 'stamp': a phonological/perceptual competitor, varying by condition; (iii) 'lorry': another perceptual competitor; (iv) grompa': a super-novel perceptual competitor. These were included to qualify the 'strength' of the lexical engagement effects.
  - Similarly to the novel competitor trials, familiar competitors were either phonological or perceptual and it was to this competitor object that fixations were measured.

How these trials, with the multi-object arrays, were adapted to mouse tracking varied between Experiments 5–7, but all the mouse tracking experiments had five trial types:

1. Novel word phonological competition trials: testing lexical engagement between a novel and a familiar object;
2. Novel word perceptual competition trials: to set a base line to contrast the phonological competition condition with;
3. Familiar word phonological competition trials: as per Spivey et al. (2005), and Experiments 3 and 4, testing familiar word lexical engagement;
4. Familiar word perceptual competition trials: again, as per Spivey et al. (2005), and Experiments 3 and 4;
5. Super-novel filler trials.

All novel objects, except for the super-novel fillers, were learnt either on, or the day before, testing, as in Weighall et al. (2017).

For several reasons, Experiments 5–7 are not presented here in chronological order. Note that the numbering of the experiments refers to the order in which they appear in this thesis, not their chronological order. Chronologically, Experiment 6 was the conducted first, in Spring 2019. It is presented as the second experiment of the set, in Chapter 13 (p. 177). It originated as a replication of Weighall et al. (2017), but mistakes were made during coding. However, the data are still presented here, as useful conclusions could nevertheless be drawn.

Experiment 5 was conducted next, to correct the mistakes of Experiment 6. It ran in the autumn and spring of the academic year 2020/21. Testing was cut slightly short by the CoViD-19 pandemic ($N = 57$, intended $N = 60$). This was a successful replication of Weighall et al. (2017). It is presented first, in Chapter 12 (p. 155), as the two other experiments deviate from this published 'standard' design (with associated consequences for the effect discussed in the relevant chapters for those experiments).

Experiment 7 is the final experiment of the thesis, presented in Chapter 14 (p. 191). It is planned to run as soon as possible, according to pandemic restrictions. Currently, there is data for only a single participant. The experiment addresses a limitation in the design of Weighall et al. (2017).

It should be noted that the reordering of Experiments 5 and 6 as presented in the thesis results in some mouse tracking methodology changes being introduced into Experiment 5, and then not appearing subsequently. This was the result of collecting the data for Experiment 6 first, and then refining the method further for the subsequent experiments. For example, an initiation time cut was introduced in Experiments 5 and 7 following a review of the recent literature on optimal mouse-tracking study procedures (e.g., Kieslich, Schoemann et al., 2020).

The key change between each of these three experiments is the novel word perceptual competition trials. In all experiments, novel word phonological competition came from a pair such as 'biscuit' and 'biscal'. Likewise, familiar word trials were maintained from Experiments 3 and 4. Super-novel filler trials compared a perceptually competing familiar object against a super-novel target, as in Weighall et al. (2017). Since they varied, the novel word perceptual competition trials for each experiment are summarised below.

11.1.1 Experiment 5 (Chapter 12): Establishing novel word competition effects

Experiment 5 was a straight and correctly implemented replication of Weighall et al. (2017), the perceptual competition trials featured a learnt novel word competitor, against a familiar word target which had not had its competitor trained (e.g., 'guitar'; 'guitas' not learnt).

11.1.2 Experiment 6 (Chapter 13): Are participants sensitive to novel word semantics in a mouse tracking task?

In their experiment, Weighall et al. (2017) refer to the condition here called 'perceptual competition' as the 'untrained' condition. With four objects in the array, Experiment 6 arose as a misunderstanding as to which object this word 'untrained' applied to. With a super-novel object in the array, it was assumed that this object was the eponymous 'untrained' object. However, as implemented in Experiment 5, the word 'untrained' actually referred to the fact that a competitor had not been trained for the familiar word *target*. Thus, the 'untrained' object from the 'untrained' condition was actually the familiar object – it was 'untrained' as no familiar competitor had been learnt.

Experiment 6's novel word perceptual competition condition compared familiar word targets, for which a competitor *had* been trained (e.g., 'biscuit', having learnt 'biscal') to a super-novel competitor. The data were still useful as, when a phonological − perceptual competition trial difference was still found, it confirmed that participants could reject a competitor object which they had not learnt more easily than a learnt competitor object in the mouse tracking task. However, this was confounded by the fact that participants were familiar with the novel competitors on phonological competition trials, which prompted the design in Experiment 7.

11.1.3 Experiment 7 (Chapter 14): The nature of novel word representations: the binding between novel words and referents

Experiment 7 built on the finding that the object was important to the effects which Weighall and colleagues had observed. From their published analyses, it is not possible to conclude that the competition which arises is driven by the novel referent, as no comparison is made between fixations to the novel competitor and fixations any other object in the array. It may have been that upon hearing a word with a stem matching the stem of a word they learnt (e.g., the 'bis–' in 'biscuit', but also in the learnt label 'biscal'), participants experienced lexical competition. Other VWP papers have demonstrated such effects, albeit, with a different design (Kapnoula et al., 2015; Kapnoula & McMurray, 2016a). In this account of processing, participants are essentially reconstructing their learning on the fly: hearing the stem and seeing an object that they learnt previously invited them to ponder how these two pieces of information are linked, as both were recognised as learnt. This is quite different, however, from the processing indicative of real word learning (particularly in the developmental literature, e.g., Carey & Bartlett, 1978; Dysart et al., 2016; Riggs et al., 2015): where learning is semantic, and a referent and a label are linked and stored as a 'multi-dimension array' (Gaskell & Marslen-Wilson, 1997). However, this processing is precluded on perceptual competition trials, as the stem of the target there was not shared by a novel word.

When it runs, Experiment 7 will address this by pairing a learnt novel referent (e.g., BISCAL) against a familiar target for whose label a novel competitor has also been learnt (e.g., if the familiar target was GUITAR, then the competitor label 'guitas' would have been learnt also). If participants have not linked the label 'biscal' to its referent BISCAL, this condition – intended to evoke perceptual competition – will evoke phonological competition, as participants wrongly apply the novel label evoked by the familiar target's label (i.e., 'guitas'), to the on-screen novel referent (i.e., the BISCAL). Experiments 5–7 follow in Chapters 12 to 14.

CHAPTER

TWELVE

EXPERIMENT 5 ESTABLISHING NOVEL WORD COMPETITION EFFECTS

12.1 Introduction and rationale

As set out in Chapter 11, Experiment 5 set out to replicate Weighall et al. (2017), and adapt the visual world paradigm (VWP; Tanenhaus et al., 1995) design to mouse tracking, applying the procedures and protocols determined in Experiments 3 and 4. It was an opportunity to attempt a replication of the novel word competition effects, and determine if mouse tracking was sensitive to such effects.

Of the effects that were reported by Weighall and colleagues, there were three of particular interest. These were that:

1. Novel words show immediate competition effects;
2. Competition effects for novel words learnt either the day before testing, or on the day of testing, were otherwise be indistinguishable;
3. That familiar words show stronger competition effects than novel words.

A difficulty in adapting VWP was the reduction of the VWP array, with four objects, down to two in a mouse tracking trial. Further comparisons between objects would need to be included as further conditions. It was decided that the experiment needed to establish competition effects for familiar and novel words (therefore, two conditions – perceptual and phonological competition trials – per word type). This allowed for comparison with Experiments 3 and 4, and represented a sort of crossing between the designs of Spivey et al. (2005) and Weighall et al. (2017). Super-novel filler trials were also included, as otherwise participants would not have clicked on any novel objects, which may have biased participants against them, eliminating potential competition effects. Further details specifying the design choices are discussed later (Section 12.2.3, p. 158).

In Experiment 5, the novel word perceptual competition trials presented a novel word which had been learnt (e.g., 'aliet'), against a target 'base' word (e.g., 'balcony', see Tables D.1 to D.3, pp. 219 and 220). Base words were words for which a

novel competitor had been constructed (but not necessarily learnt) by those participants (e.g., 'alien' → 'aliet'; 'balcony' → 'balcozo'). Crucially, for the base word presented on *perceptual* competition trials, participants had *not* learnt the derived novel competitor (e.g., having *not* learnt 'balcozo', the participant heard/saw 'balcony'). Note that the participant *had* however learnt the novel competitor appearing alongside the target, here, ALIET. Thus, when they heard the stem /bælkə—/, it could *only* evoke that base word (i.e., 'balcony'), forbidding phonological competition. However, there could still be detectable perceptual competition from the novel competitor. The difference between novel word phonological and perception competition conditions was still therefore the 'extra' competition, due to the shared phonology present in the stem.

With respect to the conclusions of Experiments 3 and 4, it was thought that path length (PL) would again show stronger effects than response time (RT). No significant differences in initiation time (IT) were predicted. Data were expected to be unimodal, also. With respect to the trajectories, no a-priori predictions were made, other than that the familiar words would show patterns similar to that seen in Experiments 3 and 4, and that these patterns would be appropriately smaller for responses driven by the more fragile novel word representations (cf., Weighall et al., 2017).

In a change from Experiments 3 and 4, it was decided that Experiment 5 should implement an IT cut, as this had been reported to give rise to larger effect sizes (Kieslich, Schoemann et al., 2020). An IT cut forced participants to initiate movement before a set initiation time; if they initiated movement after this time, a warning would appear on screen asking them to initiate faster. An appropriate value needed participants to already be in motion by the end of the carrier phrase ("Click on the..."), and thus the distribution of carrier phrases was examined. This is shown in Fig. C.2 and Table D.4 (pp. 217 and 221). The minimum carrier length was 451ms, and thus to ensure that participants were initiating movement before the end of the carrier phrase on all trials, an IT cut of 450ms was set. Although they did not impose a cut, Spivey et al. (2005) did implement 500ms of silence before participants heard their word labels – with the express purpose of ensuring participants were in motion by the time of word onset. Thus, an IT cut of 450ms was also comparable with design choices made elsewhere in the mouse tracking literature. Parameters in the mouse tracking task were otherwise unchanged from Experiments 3 and 4.

Another important question was what an appropriate sample size would be. Previous experiments in the thesis had tested samples of convenience, with recruitment being open for a set time, and not targeted. To boost the chances of replication, a more focussed recruitment strategy was implemented. However, it was difficult to estimate what would be a sufficient sample size to detect novel word competition, as the effect size for novel word competition was unknown. No estimates were available in the literature as previously only one research project had looked at novel word learning in mouse tracking, with quite a different design (Bartolotti & Marian, 2012). However, an attempt was made to estimate an appropriate size in the following way. Firstly, a power analysis for a paired-samples, two-sided t-test was conducted, with parameters of 80% power, $p \leq 0.05$ and the smallest effect size observed for *Cartoon* stimuli in Experiment 4 ($d_{RT} = 0.384$; see Table 10.4,

p. 141). This power analysis showed that sample of 56 would be sufficient to detect a difference between familiar word phonological and perceptual competition trials. As there were six counterbalancing lists, this was scaled up to a target sample size of 60 (10 participants per list).

To further boost the chances of replication, it was also anticipated that participants could be eliminated at various stages during the experiment: for example, during training, should their learning be poor (as indicated by accuracy on a two-alternative forced-choice (2-AFC) task). To ensure that the final sample, post-exclusions, was closer to 60, the decision was made to eliminate participants as the experiment progressed (see Section 12.2.4, p. 161).

12.2 Methods

12.2.1 Participants

Data were collected from 78 participants, of whom seven were excluded for not returning on a second day of testing, and 11 of whom were excluded according to procedures set out below (Section 12.2.4, p. 162). Unfortunately, due to the CoViD-19 pandemic and March 2020 lock down, testing was cut short before all 60 participants could be collected, giving a final sample of 57 participants.

Therefore, in Experiment 5, data from 57 participants were analysed (13 male, M_{Age} = 22.5 years, SD_{Age} = 6.54 years, 43 monolingual, 48 right-handed). All participants were fluent in English. Participants were all tested in a quiet laboratory environment. No participants had participated in any previous experiments. All were free of any confounding disorders (e.g., sensory, learning or language difficulties), or had corrections to normal (e.g., by wearing eyeglasses). Participants all ordinarily used the mouse with their right hand.

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, in exchange for course credits, or a small financial compensation (£16 voucher).

12.2.2 Materials and apparatus

The materials came in four blocks. The first three of these were three lists of 48 words, each with a recording, and an associated picture. A full list of these words can be seen in Tables D.1 to D.3 (pp. 219 and 220). Sample referents are seen in Figs. C.1 and D.1 (pp. 217 and 221). Materials were kindly donated by Anna Weighall, and thus Experiment 5 used the same stimuli set used in her work (Weighall et al., 2017).

Each of the first three blocks consisted of a known base word (e.g., 'angel'), and a novel competitor derived from it (e.g., 'angesh'). Each novel word was constructed by altering the base word at its uniqueness point. With respect to the three word lists, Weighall et al. (2017, pp. 4–5) reported that:

> "All base words were high frequency nouns ... with an age of acquisition of 7.5 years or less (Brown et al., 2012). The novel words were all phonemically identical to base words until the point at which the word

> [became] unique according to CELEX (M = 4 phonemes), and were created by changing the final few phonemes of the base word after the uniqueness point. Ten adults were asked to name a large pool of pictures [which] were selected [as referents for base words] if naming agreement was ≥80% .... Novel objects were included if they were not given a specific name [by the same adults], and yet were identified [by them] as belonging to one of four categories (animal, musical instrument, plant, tool). [The novel objects] were paired with the novel words using the following criteria: (1) there was no semantic overlap between the base word and the category of the novel object (e.g., the novel word 'donkop' was not paired with a novel object from the 'animal' category), and (2) there was no perceptual overlap between shape or colour properties of the base word picture and the properties of the novel word object .... The base words ... were matched [for each list] on CELEX frequency (M = 8.11, SD = 8.93), n syllables (M = 2.38, SD = 0.49), n phonemes (M = 6.35, SD = 1.08), uniqueness point (obtained from CELEX, expressed as number of phonemes from onset; M = 4.22, SD = 0.83). Novel objects and base word pictures in each list were also matched for visual complexity (including number of object features (parts) and number of colours) to ensure that novel competitor objects were not more or less salient than the base word objects."

In addition to the first three blocks, a fourth block contained 40 familiar words and 20 super-novel words. The familiar words provided a baseline to compare novel word competition against (Table D.5, p. 222). These 40 words were arranged into 20 competing pairs (e.g., 'baker', 'bacon'). Weighall et al. (2017) report that their referents again had ≥80% naming agreement amongst the norming group, and had been matched on: verbal and written frequency, concreteness, familiarity and imageability, according to the MRC Psycholinguistic Database (Wilson, 1988).

Finally, the twenty super-novel words (e.g., 'balras') in the experiment (Table D.5, p. 222), were reported by Weighall et al. (2017) as taken from the Graded Nonword Reading Test (Snowling, John, Adams, Bishop & Stothard, 2001) and the Blending Nonwords subtest of the Comprehensive Test of Phonological Processing (Wagner, Torgesen, Rashotte & Pearson, 1999).

Pictures were as in Experiment 4 (Section 10.2.2, p. 131): scaled to 300 × 300 pixels but otherwise unedited from Weighall et al. (2017), in order to preserve the norming of the set. Sound files were also as used previously in Experiment 4.

All apparatus and mouse tracking script parameters used in Experiment 5 were exactly the same as that used in Experiments 3 and 4, except for an IT cut of 450ms.

12.2.3 Design

Overview

The experiment took place over two days. The first day involved only training. The second day involved training, followed by the lexical engagement and then the lexical configuration tasks.

Upon arrival, participants were allocated to one of six experimental lists according to two variables (arranged across a 2 × 3 pattern). The first variable determined which familiar word was the target and which the competitor (i.e., for the pair CANDY and CANDLE, which of these words a participant heard). The second determined which novel words a participant would learn, and which words would contribute to the perceptual competition trials. For example, a participant on list A123 heard the word 'candle' on the CANDY–CANDLE pairing (Table D.5, p. 222), and on the first day of the experiment, learnt words from List 1 (Table D.1, p. 219). On the second day, they learnt novel words from List 2 (Table D.2, p. 220). For this participant, base words from List 3 were used in the novel word perceptual competition trials (explained later).

Words learnt on the first day of the experiment were labelled 'words learnt yesterday' ('yesterday words'). Words learnt on the second day of the experiment were labelled 'words learnt today' ('today words'). *Base* words were words which could be assumed to be known, and for which a novel competitor had been derived (whether or not the participant learnt that novel competitor). *Familiar* words were also words that were assumed to be known to the participant, and for which no competitor had been derived, as they were in a separate block of items (Lists A/B). The terms 'base' and 'familiar' here denote only the allocation to lists and if a competitor was derived: both sets contained common English words. As in previous experiments, 'perceptual' competition referred to that competition arising only from participants needing to select a response option on the left or the right of the screen, and rejecting the other object. 'Phonological' competition also referred to participants needing to make this judgement, but having the additional demand of needing to discriminate between the two objects whose labels overlapped (e.g., 'candy', 'candle'; 'angel', 'angesh').

To emphasise the continuity of the word learning processes with time, and to allow a comparison with novel words, familiar words are subsequently referred to as 'words learnt long ago', ('Long ago words'). The experiment therefore examined two variables: Competition (*Perceptual, Phonological*), and Day of Word Learning (shown below as 'Word learnt…'; *Long ago, Yesterday, Today*).

Training

The design of the training task replicated that used in Weighall et al. (2017), using computer scripts shared by Anna Weighall. Training took place in four blocks. The first three blocks required various listen and repeat tasks – Weighall and colleagues state that the purpose of this was to draw attention to the phonological form of the novel words, and that it was reflective of training used elsewhere in the literature (e.g., Brown et al., 2012; Henderson, Powell, Gaskell & Norbury, 2014). In all of these blocks, participants also saw a novel referent, which was held on screen for 2s as participants heard the novel word. The final block was a 2-AFC task, given with feedback. Across these four blocks, participants received 12 presentations of the novel words and their referents.

Lexical engagement task

Mouse tracking was used to test lexical engagement. Design of the mouse tracking task was as follows, with an example participant on list A123 receiving:

1. 10 phonological competition familiar word trials (e.g., "Click on the candle"; CANDY and CANDLE present);

2. 10 perceptual competition familiar word trials (e.g., "Click on the sandal"; SANDAL and BAKER present);

3. 12 phonological competition novel word trials, for yesterday words (List 1 objects; e.g., "Click on the alien"; ALIEN and ALIET present);

4. 12 perceptual competition novel word trials, for yesterday words (a List 1 novel competitor against a List 3 base word; e.g., "Click on the athlete"; ATHLETE and GRAFFINO present);

5. 12 phonological competition novel word trials, for today words (List 2 objects; e.g., "Click on the angel"; ANGEL and ANGESH present);

6. 12 perceptual competition novel word trials, for today words (a List 2 novel competitor against a List 3 base word; e.g., "Click on the mushroom"; MUSHROOM and WALRICK present);

7. 20 super-novel filler trials (e.g., "Click on the balras"; BALRAS and KETTLE present)

The numbers of trials directly followed Weighall et al. (2017). Within each group of trials given above, half had a target placed on the right, and half placed on the left. Note that the familiar items appearing on the super-novel filler trials were not used elsewhere in the experiment. No item appeared more than once. Preceding the experimental trials were 16 practice trials, using different stimuli and words, taken from Experiment 3, in accordance with mouse tracking best practice (Kieslich, Schoemann et al., 2020). Except for these practice trials, which were blocked to occur before the experiment, all other trials occurred interleaved.

Lexical configuration tasks

Two cued-recall tasks were used to assess lexical configuration. These tasks allowed for the assessment of explicit knowledge of novel object's form and referent. The first task, stem completion, required only knowledge of the phonological form of a novel word, whereas the second (picture naming) required knowledge of the referent as well. The percentage of items correctly recalled from each of the two days of training was recorded, and compared across tasks and days. RT was not reported, following Weighall et al. (2017).

12.2.4 Procedure

Participants began on the first day of the experiment with a training task, which had four blocks, as follows. Training was conducted in E-Prime (Schneider, Eschman & Zuccolotto, 2002). Each block was repeated twice, with items in a random order. Blocks 1–3 occurred with the novel object held on screen for 2s, and half a second between trials.

1. Block 1 – whole word listen and repeat task. Participants heard the whole word, and were told to repeat back the whole word (e.g., heard 'baboop', said 'baboop').

2. Block 2 – initial segmentation listen and repeat task. Participants heard the whole word, and were told to repeat back the first syllable (e.g., heard 'baboop', said /bæ/).

3. Block 3 – final segmentation listen and repeat task. Participants heard the whole word, and were told to repeat back the final syllable (e.g., heard 'baboop', said /buːp/).

4. Block 4 – 2-AFC, with feedback. Participants saw on screen, fixed for 2s, two novel referents they had learnt in the previous blocks, and heard a novel word. Participants had to press either '1' or '9' on the keyboard to indicate whether the novel word had previously been associated with an object on the left, or right, of the screen, respectively. Regardless of their response, the novel object then appeared on screen alone for a further 2s and participants again heard the novel word, as feedback. No response was required here, and the task then progressed to the next trial. Within this block, there were 72 trials (three trials × 24 items).

Participants therefore heard the novel words and saw the novel objects 12 times – twice in each of Blocks 1–3, and then twice on each 2-AFC trial (once on the trial itself, once during feedback), then repeated three times. Having completed training in a little under 30 minutes, participants left, and returned the next day.

On the second day of the experiment, taking place exactly 24 hours after the first session, participants performed training again, with new words according to their list, as explained above (Section 12.2.3, p. 158). Following this, they immediately performed the lexical engagement task (mouse tracking). The procedure for this was identical to that described in previous chapters, with the design of the task also given above. The only change was the introduction of the IT cut. If participants did not initiate movement in under 450ms, a warning appeared on screen instructing participants to initiate movement faster, even if they were not entirely sure of a response yet (cf., Freeman & Ambady, 2010; Kieslich, Schoemann et al., 2020). Participants were not explicitly told to watch for this, but became accustomed to it during the practice trials. Participants were offered the opportunity to repeat the practice trials if they felt necessary; none did.

Participants finished testing with the lexical configuration tasks. Response accuracy was recorded by pen and paper. Both tasks took place in the same order.

The first lexical configuration task was stem completion. Here, participants heard the novel word, cut at the disambiguation point (e.g., /bæbuː—/). At entirely their own pace, participants had to produce the correct novel form (i.e., 'baboop', /bæbuːp/). Words from each day of testing were interleaved in a random order, and participants could hear each stem as many times as they liked.

The second lexical configuration task was picture naming. Here, participants saw a novel referent on screen, and had to produce the form that had been associated with it. Again, trials were interleaved across lists, and participants proceeded at their own pace.

In both lexical configuration tasks, to allow for differences in accent etc., responses with incorrect vowels were accepted as correct if the consonants were correct. For example, whilst during training participants heard 'rugbock' pronounced with a southern English vowel (/ɹʌgbɐk/), this was frequently realised as /ɹʊgbɐk/ by northern speakers at test. This deviation was accepted.

Processing of data and exclusions

Processing of data took place as in previous experiments (see Section 9.2.4, p. 114). Additionally, trials were filtered out according to the IT cut (see below). Data were taken on the measures identified in Experiment 4: IT, PL and RT. The distributions and trajectories were also examined as previously.

Training. Exclusions were planned for participants with $< 75\%$ accuracy on the final 2-AFC task, separately for each day. However, this did not apply to any participants (minimum performance was 76.4%).

Lexical engagement task. Trials were resampled, time and space normalised, and remapped to the right side of the screen. Practice and super-novel filler trials were removed, and data from the 11 participants that had been identified as inaccurate responders were not analysed further. Inaccurate responders had been identified during testing, but before analysis, by examining the raw error counts before data processing. Any participant with $< 70\%$ accuracy on any of the design cells set out above (Section 12.2.3, p. 158) was not analysed further. This left 57 participants, whose 3876 (68×57) trials were processed as follows:

- 76 incorrect responses (1.96% of trials) removed;
- 39 trials with an IT longer than the cut (450ms) removed;
- 121 trials removed by a $M \pm 3SD$ trim on PL and RT (as previously, per condition);
- 251 trials removed by excluding a further 5 participants, for having less than 70% of their trials remaining in any design cell.

This left 998 *Long ago*, 1199 *Yesterday* and 1192 *Today* word trials. On average, the remaining 52 participants contributed 65.2 trials out of a maximum of 68 (95.8%) for analysis.

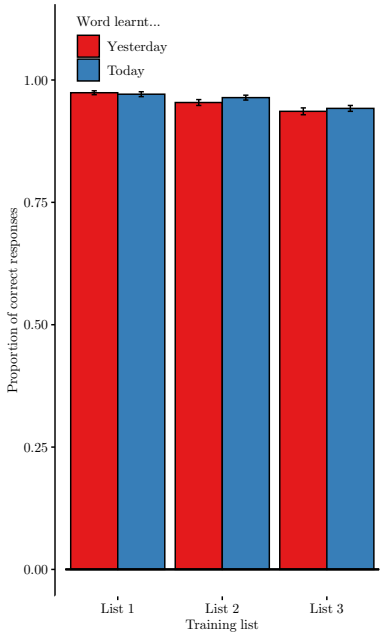


Figure 12.1: Training task (2-AFC) performance in Experiment 5. Error bars show ±1 *SE*

Lexical configuration task. Participants who were excluded from the lexical engagement task during analysis were still allowed to contribute to the lexical configuration dataset, as they had shown sufficient (i.e., $\geq$ 75% accuracy) training performance. However, due to a computer crash on the second day of testing, data from a single participant who had been included in the lexical engagement task had to be excluded. Therefore, 56 participants contributed lexical configuration data.

12.3 Results

Mouse tracking data were collected in MouseTracker (Freeman & Ambady, 2010). All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016), and the mouse tracking data were processed with `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff et al., 2020).

12.3.1 Training

Only data from the final block of training, the 2-AFC task, were analysed, as the listen and repeat blocks did not assess participants' knowledge of the items, only their ability to echo their forms. On average, across all days and lists, participants paired the correct referent to the heard label 95.7% (SD = 4.56%) of the time. Average recognition performance across day and lists can be seen in Fig. 12.1 (p. 163).

Two ANOVAs, one for each day, were performed to test for equal list performance on each day. Words learnt on both days showed equivalent performance across lists (words learnt before testing: $F(2, 54) = 2.63$, $p = 0.081$, *NS*, $\eta_g^2 = 0.089$; words learnt

Table 12.1 Summary of descriptive statistics for each lexical engagement measure, by when the word was learnt, and competition, in Experiment 5

Word learnt...	**Measure**	**Competition**			
		Perceptual		*Phonological*	
		M	*SD*	*M*	*SD*
Long ago	*IT (ms)*	88	51	87	51
	PL	1.98	0.212	2.29	0.348
	RT (ms)	1593	138	1654	123
Yesterday	*IT (ms)*	90	52	84	47
	PL	2.01	0.245	2.07	0.257
	RT (ms)	1644	139	1722	170
Today	*IT (ms)*	86	52	88	46
	PL	1.97	0.195	2.05	0.212
	RT (ms)	1631	139	1665	140

Note. PL units are arbitrary.

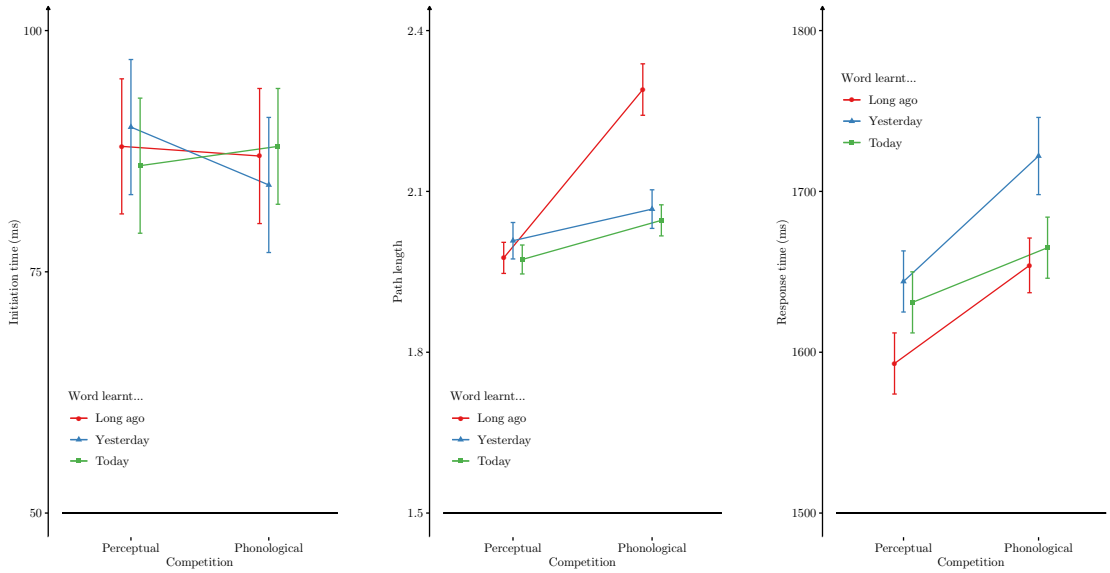
on the day of testing: $F(2, 54) = 3.08$, $p = 0.054$, *NS*, $\eta_g^2 = 0.101$)[1]. Therefore, all further analyses collapsed across training lists.

12.3.2 Lexical engagement

Overview of the mouse tracking measures

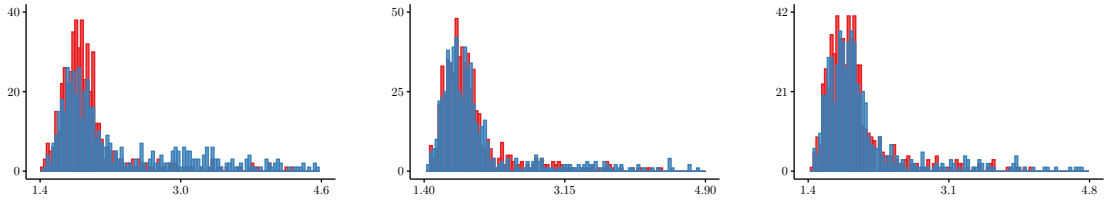
Each mouse tracking measure is presented below in turn. Descriptive statistics for all measures are seen in Table 12.1 (p. 164) and Fig. 12.2 (p. 165). The plan for the analysis was first to compare novel word trials in a 2 × 2 Competition (*Perceptual, Phonological*) by Day of Word Learning (*Yesterday, Today*) repeated-measures ANOVA for each measure (summarised in Table 12.2, p. 166). A further ANOVA was then conducted (see Table 12.3, p. 167), inputting the same variables but also including words learnt long ago. For the second ANOVA (comparing novel and familiar words), the plan was to collapse across day of word learning if in the first ANOVA (comparing novel words only) there was no main effect of day, nor an interaction between day and competition. Post-hoc *t*-tests were then performed as appropriate (see Table 12.4, p. 168). The distribution of the data and the trajectories were also analysed.

[1]Note that day of word learning could not be entered into the ANOVA as the rotation across lists was not the same for all participants, e.g., participant 1 first learnt words on List 1, then learnt the words on List 2; however, participant 41 first learnt words on List 3, then on List 1



(a) IT (b) PL (c) RT

Figure 12.2: Means' plot for each mouse tracking measure in Experiment 5. Error bars show ±1 *SE*



(a) Long ago (b) Yesterday (c) Today

Figure 12.3: Histograms for PL by word learning time in Experiment 5. Perceptual competition trials shown in red; phonological competition trials shown in blue

Table 12.2 Summary of the day by competition ANOVAs for novel word trials for each of the mouse tracking measures in Experiment 5

Measure	Effect	F	p	η_g^2
IT	*Day*	0.013	0.908, *NS*	< 0.001
	Competition	1.17	0.285, *NS*	< 0.001
	Day × *Competition*	1.43	0.237, *NS*	0.002
PL	*Day*	1.48	0.229, *NS*	0.004
	Competition	12.1	0.001^{***}	0.021
	Day × *Competition*	0.127	0.723, *NS*	< 0.001
RT	*Day*	11.5	0.001^{***}	0.014
	Competition	21.6	0.001^{***}	0.036
	Day × *Competition*	5.68	0.021^{*}	0.006

Note. $df = (1, 51)$. Three asterisks (***) denotes significance at the 0.001 level. A single asterisk (*) denotes significance below the 0.05 level.

Distributional analysis

Spatial responding to all types of words was unimodal, according to Hartigans' dip statistic (all ps ≥ 0.543, *NS*). This was confirmed by visual inspection of the histograms (see Fig. 12.3, p. 165). The data appeared to again be skewed, but this was not considered a problem given the size of the sample and the tests that were performed (Blanca et al., 2017; Ghasemi & Zahediasl, 2012; Lumley et al., 2002).

Initiation time

Descriptive statistics. As seen in Table 12.1 and Fig. 12.2 (pp. 164 and 165), ITs were all very similar across conditions (84–90ms), and much faster than in previous experiments.

Yesterday/Today novel word comparisons. As seen in Table 12.2 (p. 166), the novel word ANOVA showed no main effects, and no interaction (all ps ≥ 0.237). Novel word trials were therefore collapsed across day of learning for comparison with familiar words.

Collapsed novel and familiar word comparisons. As seen in Table 12.3 (p. 167), the collapsed novel and familiar word ANOVA also showed no main effects, or an interaction (all ps ≥ 0.438). This confirmed that differences in the other measures were not due to selectively different ITs across conditions.

Table 12.3 Summary of the day by competition ANOVAs, for words learnt long ago and recently, in Experiment 5, for each of the mouse tracking measures

Measure	**Effect**	F	p	η_g^2
IT	*Day*	0.049	0.826, *NS*	< 0.001
	Competition	0.612	0.438, *NS*	< 0.001
	Competition × *Word*	0.014	0.908, *NS*	< 0.001
PL	*Day*	21.9	$< 0.001^{***}$	0.047
	Competition	78.8	$< 0.001^{***}$	0.131
	Day × *Competition*	27.2	$< 0.001^{***}$	0.060
RT	*Day*	41.2	0.001^{***}	0.040
	Competition	16.2	0.001^{***}	0.029
	Day × *Competition*	2.47	0.090, *NS*	0.004

Note. $df = (1, 51)$ for IT and PL effects. RT day $df = (1, 51)$; other RT effects $df = (2, 102)$. IT and PL data were collapsed over words learnt yesterday and today, but the comparisons for RT were not, due to those main effects shown in Table 12.2 (p. 166). Three asterisks (***) denotes significance at the 0.001 level.

Table 12.4 Summary of competition effect *t*-tests for each type of word, for PL and RT, in Experiment 5

Measure	Word learnt...	*t*	*p*	*d*
PL	*Long ago*	7.67	$< 0.001^*$	1.03
	Recently†	3.47	0.001^*	0.332
RT	*Long ago*	4.62	$< 0.001^*$	0.461
	Yesterday	4.62	$< 0.001^*$	0.491
	Today	2.60	0.012^*	0.247

Note. $df = 51$. An asterisk (*) denotes significance $\alpha = 0.017$, due to the Bonferroni correction. † indicates that novel words learnt yesterday and today are collapsed (see Table 12.2, p. 166).

Path length

Descriptive statistics. As seen in Table 12.1 and Fig. 12.2 (pp. 164 and 165), descriptive statistics suggested a competition effect for familiar and novel words. The size of the effect appeared to be larger for familiar words. Whilst the size of the effect was similar for novel words learnt yesterday (0.06 units) and today (0.08 units), there was more variability in responses to novel words learnt yesterday, as indicated by larger *SD*s ($\sim$ 0.25 units, compared to $\sim$ 0.2 units).

Yesterday/Today novel word comparisons. As seen in Table 12.2 (p. 166), whilst a significant effect of competition was observed ($p < 0.001$), there was no significant main effect of day of learning ($p = 0.229$), nor an interaction ($p = 0.723$). Therefore, for comparison with familiar words, the novel word trials were collapsed across day of learning.

Collapsed novel and familiar word comparisons. As seen in Table 12.3 (p. 167), main effects of competition and day of learning and an interaction were observed (all *p*s < 0.001).

Post-hoc *t*-tests. Further comparisons were conducted to determine if both sets of words showed competition effects, and the size of these effects. As seen in Table 12.4 (p. 168), this showed that both words learnt long ago and words learnt recently engaged their competitors (all *p*s ≤ 0.001). The effect size for words learnt long ago was comparable to that observed in Experiment 4 ($d_{Expt4} = 0.988$, $d_{Expt5} = 1.03$), and was much larger than that for novel words ($d_{novel} = 0.332$). The much larger competition effect seen on familiar word trials compared to novel word trials therefore explains the interaction observed between competition and day.

Response time

Descriptive statistics. Table 12.1 and Fig. 12.2 (pp. 164 and 165) shows that responses on novel word trials were slower than responses on familiar word trials for both perceptual and phonological competition. Responses on yesterday word trials were slower than responses on today word trials.

Yesterday/Today novel word comparisons. As seen in Table 12.2 (p. 166), RT showed main effects of competition, day of learning, and an interaction. This suggested that responses on yesterday word trials were significantly slower than on today word trials, but that this varied across the competition trials. Unlike PL, RT trials were therefore not collapsed over day of word learning.

Novel and familiar word comparisons. A 2 × 3 repeated-measures ANOVA was performed (see Table 12.3, p. 167), comparing competition for all three levels of day of word learning. This again showed main effects for competition and day of word learning (both ps < 0.001), but no interaction ($p = 0.090$).

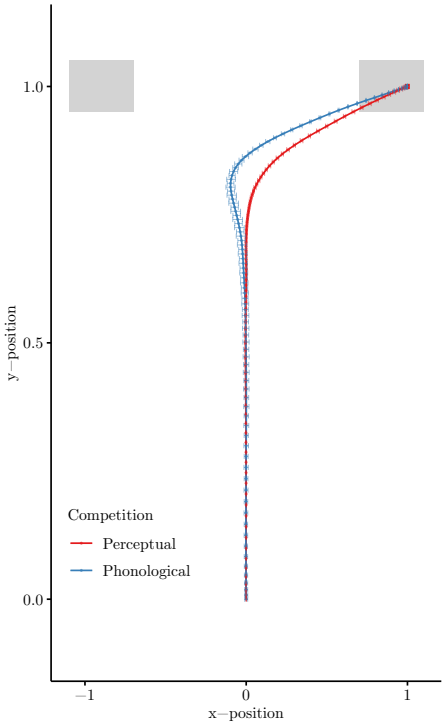
Post hoc *t*-tests. The RT data competition effects were observed for words learnt at all time points (see Table 12.4, p. 168). Words learnt yesterday and long ago showed very similar effect sizes (both ps < 0.01, $d_{Long\ ago} = 0.461$, $d_{Yesterday} = 0.491$). Words learnt today however showed a weaker competition effect ($p = 0.012$, $d = 0.247$). This weaker competition effect explains the interaction observed between competition and day of word learning.

Trajectory analysis

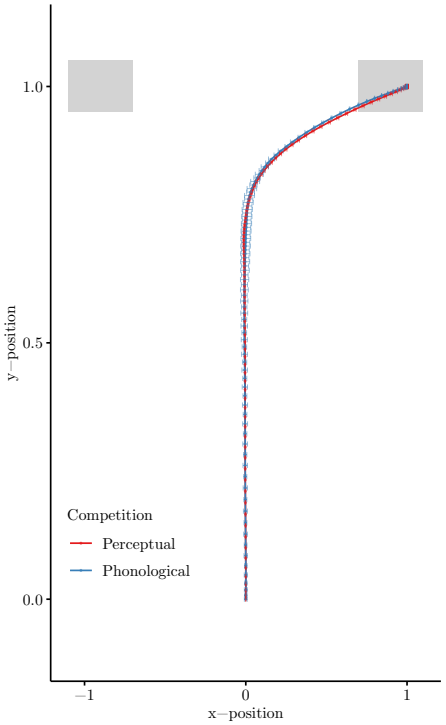
Averaged trial trajectories. Trajectories by when a word was learnt are shown in Fig. 12.4 (p. 170). Trajectories for words learnt long ago, and for words learnt today, showed additional deflection towards the competitor on phonological competition trials, whereas trajectories for words learnt yesterday did not appear to demonstrate such a difference. Newly acquired words appeared to show weaker competition effects than words learnt ago. Also noticeable, relative to the data from Experiments 3 and 4, was the much straighter vertical motion, and then a visible 'bump' of competition towards the phonological competitor for words learnt long ago.

Standardised time bin analysis. As in Experiment 4, unimodal data in Experiment 5 allowed for the comparison of x-position at each of 101 time bins for all words. Point-to-point analyses were carried out at each standardised time bin to look for differences in the x-position across perceptual and phonological competition trials. As before, a run was only accepted if it consisted of more than eight bins, to control for multiple comparison (Dale et al., 2007). The trajectories are shown in Fig. 12.5 (p. 171).

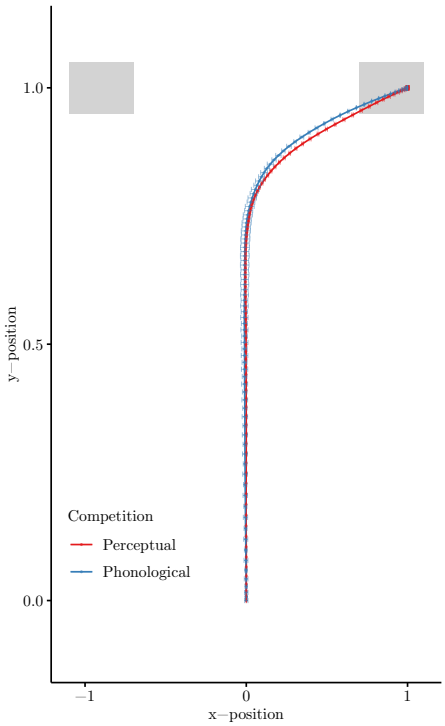
- **Words learnt long ago.** For words learnt long ago, a significant difference in x-position was observed across competition types from the 48$^{\text{th}}$ to the 89$^{\text{th}}$ time



(a) Words learnt long ago

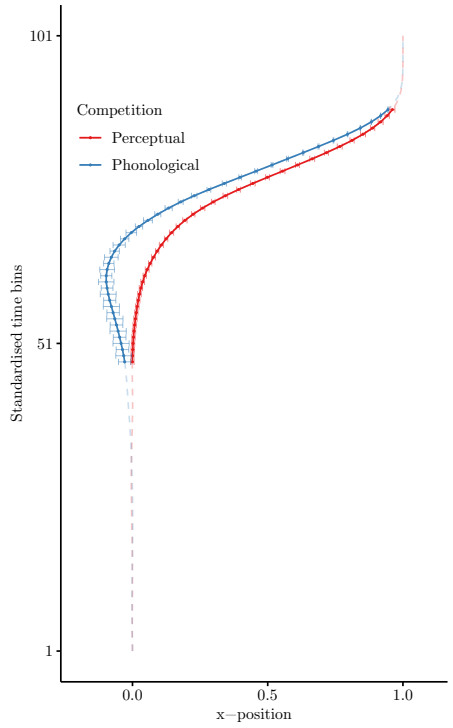


(b) Words learnt yesterday

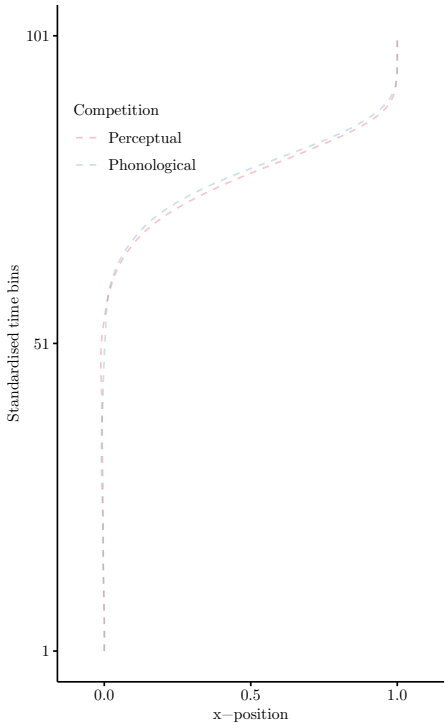


(c) Words learnt today

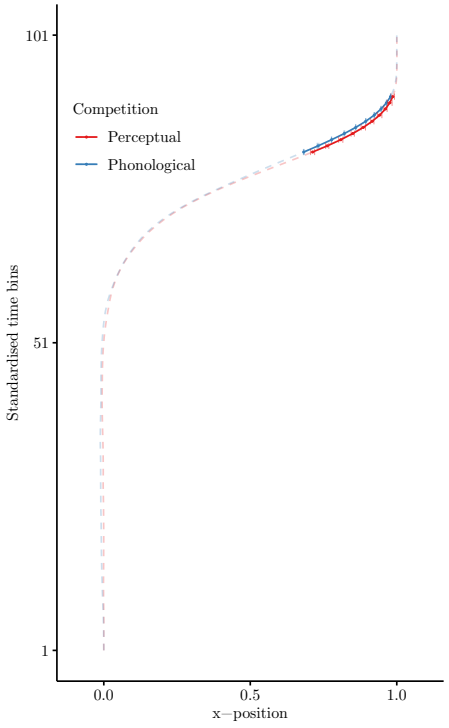
Figure 12.4: Averaged participant trajectories plotting x- against y-position for each time of word learning in Experiment 5. Each point represents a single time bin. Error bars represent $\pm$ 1SE in the x-position



(a) Words learnt long ago



(b) Words learnt yesterday



(c) Words learnt today

Figure 12.5: Averaged x-position against standardised time in Experiment 5, for words learnt at each time point. The solid saturated line (if present) indicates significantly different positions across conditions in each time bin (paired-samples t-test; $p < 0.05$). Each point represents a single time bin. Error bars represent $\pm$ $1SE$ in the x-position

bin – comparable to Experiment 4 (see Fig. 12.5a, p. 171; compare Fig. 10.5, p. 142).

- **Words learnt yesterday.** The *t*-tests for words learnt yesterday (Fig. 12.5b, p. 171) showed differences in position across competition trial types from the 86th to 91st time bins. However, as the bins did not occur in a run of eight bins or greater, this could not be said to be significant (cf., Dale et al., 2007).

- **Words learnt today.** The *t*-tests for words learnt today (Fig. 12.5c, p. 171) showed significant differences in position between the 82nd and 91st time bins.

12.3.3 Lexical configuration

Lexical configuration was assessed by two cued-recall tasks: stem completion and picture naming. Both required the verbal production of the novel word's form, however, they differed in how this was cued. Stem completion cued participants with the novel words, less the portion after the disambiguation point. Picture naming provided participants with the referent object.

In addition to analysing how performance varied across time, since it was expected that words learnt yesterday would show consolidation effects, performance across tasks was also examined. This was interesting as the two tasks allowed one to look at semantics and phonology separately, even if both tasks required a response which required only production of the form (i.e., rather than recall of explicitly semantic information, e.g., Dumay et al., 2004).

Descriptive statistics

Stem completion performance for words learnt yesterday was 53.6% (SD = 21.8), but only 29.2% (SD = 15.3%) for words learnt today. Picture naming performance was more or less flat: 20.5% (SD = 18.4%) for words learnt yesterday, and 21.6% (SD = 18.0%) for words learnt today.

Inferential statistics

A 2 × 2 repeated-measures ANOVA was conducted, inputting day of word learning (yesterday, today) and task (stem completion, picture naming). There was a main effect of day of learning ($F(1, 55) = 27.8$, $p < 0.001$, $\eta_g^2 = 0.091$), and also task ($F(1, 55) = 153$, $p < 0.001$, $\eta_g^2 = 0.235$). There was also a significant interaction ($F(1, 55) = 123$, $p < 0.001$, $\eta_g^2 = 0.108$).

Post-hoc *t*-tests showed that this came about as a result of flat performance across days for the picture naming task ($t = -0.539$, $p = 0.592$, NS, $d = 0.061$), but evidence of consolidation in the stem completion task, as overall performance was better for words learnt yesterday ($t = 8.57$, $p < 0.001$, $d = 1.27$). Furthermore, the *t*-tests confirmed that the novel word was easier to produce when cued with a stem, rather than the referent, for words learnt yesterday ($t = 13.9$, $p < 0.001$, $d = 1.62$), and today ($t = 4.94$, $p < 0.001$, $d = 0.444$).

12.4 Discussion

Experiment 5 successfully replicated Weighall et al. (2017) on all of the main findings: there was evidence of immediate competition between novel and base words, and the novel words were otherwise statistically indistinguishable from each other on the PL data in the lexical engagement task. However, the RT data did suggest that words learnt yesterday and today differed. Finally, words learnt long ago, compared to novel words, gave rise to stronger effects.

One surprising finding was the observation that when measured by RT, the effect size of the competition effect for words learnt yesterday was stronger than that for words learnt today. The size of the effect for words learnt yesterday appeared equivalent to that for familiar words, if not slightly larger. As the lexical configuration data showed very large consolidation effects (stem completion performance increased by approximately 25% after a night of sleep), it is possible that this is related to consolidation. Perhaps because participants recalled the words from yesterday better, those words were more salient, and therefore, competition – at least when measured by RT – was greater. However, why the associated d_{RT} for words learnt yesterday was a small amount larger than that for words learnt long ago is unclear. Further, this pattern was not observed in the PL data. Here there was no evidence that the competition effect was stronger for words learnt yesterday compared to words learnt today. Instead, both sets of novel words appeared to show weaker competition effects compared to words learnt long ago. It may simply be that RT is a more noisy and less reliable measure, giving rise to odd patterns (cf., Maldonado et al., 2019). Compared to PL, RT has given rise to weaker competition effect sizes for familiar words in all Experiments 3–5 (suggesting it may indeed be a less sensitive).

12.4.1 'Lexical' competition?

In their paper, Weighall et al. (2017) concluded that despite exhibiting competition effects, the novel words in their study were not fully/truly 'lexical', insofar as they exhibited a qualitatively different response profile from known words (cf., Fig. 11.1, p. 150). Also, they suggested that there was still support for abstractionist, complementary learning systems accounts of word learning – previously argued to bring about significant behavioural change after a *single* night of sleep (Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010) – as their recall data showed evidence of consolidation (i.e., an overnight improvement in the proportion of forms recalled, suggesting some stabilisation of the representation). What is not clearly articulated is why this finding is not mirrored in the lexical engagement task. Data from the replication herein also gave a mixed picture – whilst the stem completion and RT data seem to show evidence of consolidation, the PL and picture naming data did not. Indeed, the time bin analysis of the trajectories suggested that competition was not present for words learnt yesterday.

Irrespective of the conflicting evidence regarding the patterns of consolidation, what is consistent across the data from Weighall and colleagues, and from both the PL and RT data reported in the replication here, is that competition effects for novel

words do emerge immediately after learning. Therefore, the next logical step is to consider whether this competition is lexical. A limitation of Experiment 5 is that a conclusion to this question cannot be reached solely on the above data – further data are needed.

This consideration necessitates a reflection on what the end point of consolidation is, and what 'lexicality' means. From a complementary learning systems perspective, lexicality means a representation is 'non-episodic': 'abstract' and 'generalised'. This account would argue that an episodic form undergoes consolidation (a mechanism for 'lexicalisation'); the end point of processing is when that form is consolidated into the lexicon. However, a functional definition is freer of assumptions, and lexicality may instead be defined by items conforming to the predictions of speech perception models. From this perspective, lexical items are simply those which exhibit lexical properties, such as the ability to engage in competition (e.g., Kapnoula & Samuel, 2019). Whether a trace is episodic, or not, is irrelevant. Furthermore, there is no end point of processing, as stored representations may be continually updated (cf., Cai et al., 2017; Rodd et al., 2016). Much data show that word traces may contain episodic indexical information (Cai et al., 2017; Goldinger, 1996, 1998; Kapnoula & McMurray, 2016b; Kapnoula & Samuel, 2019; Rodd et al., 2016) – given that these word traces also engage in lexical engagement, it would be odd not to call them lexical. This view is inconsistent with abstractionist accounts of the lexicon (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland & Elman, 1986), but consistent with episodic accounts (e.g., Goldinger, 1998). One possibility is that the effects in Experiment 5 were driven by such episodic representations – of the sort which are thought to underpin recognition accuracy shortly after training (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995).

A parallel question is how competition arises, and the content of these novel word representations. One possibility is that 'multi-dimensional arrays', pertaining to the episode, are stored (cf, Gaskell & Marslen-Wilson, 1997). This account is simple: the label and referent are bound and stored in a representation which may then interfere with other representations. Another possibility is that the episodic representations are isolated and fragmentary – this would explain why participant were so much worse at picture naming. Fragments of the episode could have been recombined on-the-fly to produce a competition effect approximating 'lexical' competition. Responding to the on-screen objects on phonological competition trials in Experiment 5, participants heard a word stem that was shared by the competitor and the target. In the case of novel word trials, the picture of a recently-learnt competitor was also present. The 2-AFC and cued recall data suggested that participants had no difficulty recognising these trained referents. Recognising both the stem, and the picture, a participant may have been 'cued' into competition after correctly inferring that these two pieces of information had been previously paired. Note that this is the 'on-the-fly' part of processing – participants are *not*, by this account, doing anything more than recognising the novel referents and stems of novel words. Even without having a unified, 'multidimensional array' representation of the novel word, a participant could have reconstructed the novel word from the information present in the task. This recombination would then make responding to the target more inefficient, approximating 'lexical competition'. By contrast, on

perceptual trials the target did not evoke a learnt novel word (as the stem was not shared), allowing participants to respond efficiently.

Questions about the lexicality of the competition effect are addressed further in Experiments 6 and 7. Whilst it originated as a coding error, Experiment 6 was able to provide further insight, and is next presented, in Chapter 13.

CHAPTER

THIRTEEN

EXPERIMENT 6 ARE PARTICIPANTS SENSITIVE TO NOVEL WORD SEMANTICS IN A MOUSE TRACKING TASK?

13.1 Introduction and rationale

The discussion of Experiment 5 (Section 12.4, p. 173) questioned whether the competition effects observed between novel and familiar words were driven by the interplay and activation between phonology and semantics, or merely by the recognition of a shared novel/familiar word stem. Experiment 6 went some way to addressing this by testing participants' sensitivity to the novel word semantics.

The distributed cohort model of speech perception (DCM; Gaskell & Marslen-Wilson, 1997) predicts that a 'cohort' of activated representations will compete for activation (cf., Fig. 2.2, p. 11). Although the model suggests that semantics are important, and are not divorced from the word recognition process (Gow & Olson, 2015; Spivey, 2016), many word learning experiments have not trained semantics (e.g., Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Lindsay et al., 2012; Lindsay & Gaskell, 2013), and semantics are not a prerequisite for a lexical competition effect (Dumay et al., 2004; Henderson et al., 2013; Kapnoula et al., 2015; Kapnoula & McMurray, 2016a). Recent work has shown that novel words may immediately be evoked by phonological cues (such as a shared stem), without the involvement of any semantics (e.g., Fernandes et al., 2009; Kapnoula et al., 2015; Kapnoula & McMurray, 2016a). The question therefore remains if semantics supported the competition effect observed in Experiment 5 (Chapter 12, p. 155).

Knowing the content of the lexical representations in Experiment 5 is an important extension to its findings. Note that in Experiment 5 the phonological and perceptual competition conditions differed according to whether the target base word shared its stem with a newly trained phonological competitor, or not. One possibility was that participants experienced competition because of that difference alone: in the phonological competition condition the newly learnt novel words may have been evoked by the shared stem only (cf., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a). This evoking of a novel word could have interfered with the processing of the target base word, resulting in the longer PL in the phonological

competition condition relative to the perceptual competition condition. By this account of participants' processing, they were not influenced by whether (or not) the novel competitor objects on the screen were referents for that evoked novel word. However, the 2-AFC task at the end of training confirms that the representation was at that point semantic. For there to be no reference to the referent during lexical processing, this would in turn suggest either a rapid degrading of the representation (i.e., losing the semantic information – perhaps explaining the poor picture naming performance), or an inability to access this semantic information during lexical processing, as in the mouse tracking task. Either way, such non-semantic representations driving a competition effect would suggest that those representations were not fully 'lexical' in the same way as familiar words. This would be interesting, as it would suggest phonological and semantic information is handled by the cognitive system in quite different ways, with different processing time courses during learning. Indeed, this idea is hinted at by data in the literature showing that semantic effects are slower to emerge (e.g., Coutanche & Thompson-Schill, 2014; Dumay et al., 2004).

Thus, the content of the representations, and investigating the role of semantics in word learning, was quite critical. In the first instance, it was an important extension to scientific knowledge, and not otherwise addressed in the literature (cf., Weighall et al., 2017). In the second, whether a novel representation was semantic was an important qualifier of its 'lexical' status immediately after learning.

13.1.1 Addressing the semanticity of novel word representations

Experiment 6 arose as a misinterpretation of Weighall et al. (2017)'s usage of the word 'untrained' (Section 11.1.2, p. 152). On novel word perceptual competition trials Weighall et al. compared base words for which a competitor was *not* learnt (e.g., 'athlete', participant not learning the novel competitor 'athlove') with a learnt novel competitor object (e.g., LANTOBE, a novel competitor for 'lantern'). However, this misinterpretation meant that the perceptual competition trials of Experiment 6 compared a base word for which a novel competitor *was* learnt (e.g., 'alien', having learnt 'aliet') to a super-novel object (un-named). The phonological competition trials still compared base words and their learnt novel competitors (e.g., APRICOT and APRICAM on-screen). Therefore, in Experiment 6, in both conditions there was the potential for the target base word to evoke a newly learnt novel word, which was not the case in Experiment 5. In Experiment 6, the conditions differed according to whether the on-screen competitor object was a referent for that evoked novel word, or not – not by whether a novel word was evoked, or not. Such a design would allow one to test the above explanation of the competition effect observed in Experiment 5.

Observing a difference between perceptual and phonological novel competition trials in Experiment 6 would suggest one of two things. The first was that phonology alone was *not* evoking the novel word, and therefore a role for processing of the referent in the lexical engagement task. The second was that the competition effect was driven largely by phonology (cf., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a), but concurrently, participants correctly recognised that the super-novel

object was not one that appeared during training, and that it therefore could not be the referent for the evoked novel word. Thus, participants would deviate towards it less, giving the competition effect. Either of these explanations would be evidence of a truly 'lexical' and semantic representation supporting the competition effect. This was not possible to conclude from Experiment 5, as the perceptual competition condition there did not allow the *possibility* of phonological competition.

By contrast, if in Experiment 6 no competition effect were to be observed, this would suggest that the competition effect observed in Experiment 5 was the result of phonological processing alone, and that in turn the semantic information during training was for some reason inaccessible to the participants. This would imply that the representation was not lexical (see above).

The confound of familiarity

Since Experiment 6 was undertaken with the misunderstanding above, the design is not optimal. The obvious confound is that the competitor objects on the perceptual competition trials were super-novel, whereas the competitor objects on the phonological competitor trials had been familiarised during training. This difference in familiarity might also have implied differences in saliency. However, as a first attempt at investigating the role of semantics in word learning, it is presented below. The confound is addressed and resolved in Experiment 7.

Initiation time cut

Another issue with Experiment 6 was that it was run chronologically before Experiment 5. Experiment 5 design and protocols are therefore a refined version of Experiment 6. One consequence of this is that unlike Experiment 5, Experiment 6 did not implement an initiation time (IT) cut, contrary to mouse tracking best practice (e.g., Kieslich, Schoemann et al., 2020). This forced participants to initiate movement before a set value – in Experiment 6, 450ms (see Table D.4 and Fig. C.2, pp. 217 and 221). The IT cut was re-implemented in Experiment 7.

13.2 Methods

13.2.1 Participants

Sixty participants contributed data (nine male, $M_{Age} = 21.8$ years, $SD_{Age} = 5.2$ years, 43 monolingual, 51 right-handed). This sample size had been set according to the logic set out in Section 12.1 (p. 155)[1]. Although the CoViD-19 pandemic affected data collection for Experiments 5 and 7, as Experiment 6 was run before these experiments, it was unaffected. All participants were fluent in English and were tested in a quiet laboratory environment. No participants had participated in any previous experiments. All were free of any confounding disorders (e.g., sensory,

[1]Unlike in Experiment 5, in Experiment 6, participants were not rejected during data collection. This was another, later refinement to procedure introduced to Experiment 5

learning or language difficulties), or had corrections to normal (e.g., by wearing eyeglasses). They all ordinarily used the mouse with their right hand.

Participants were all tested according to procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. Participants volunteered their time freely, in exchange for course credits, or a small financial compensation (£16 voucher).

13.2.2 Materials and apparatus

Materials and apparatus were exactly as in Experiment 5. No changes were made, except for the missing IT cut in the mouse tracking script. For a full list, see Tables D.1 to D.3 and D.5 (pp. 219–222).

13.2.3 Design

As in Experiment 5, Experiment 6 compared novel words learnt yesterday, or today, to words learnt long ago. Trials were also organised by *Competition* – perceptual, or phonological. Structurally, the experiment was exactly the same, with identical Experiment 5 training and lexical configuration tasks at identical time points.

The only difference was the novel word perceptual competition trials. On these trials, objects from the list which a participant did not learn were placed against familiar base word targets for which a novel competitor had been learnt. For example, for participants on list A123, a novel word learnt yesterday perceptual trial would compare a List 1 base word target (see Table D.1, p. 219, e.g., 'alien'), for which the participant would have learnt a novel competitor (e.g., 'aliet') against an *unlearnt* List 3 novel object (see Table D.3, p. 220, e.g., ATHLOVE). For novel word learnt today perceptual trials, a List 3 novel object would also have been used, but this time, against a List 2 base word, which, again, had had a novel competitor trained (see Table D.2, p. 220).

13.2.4 Procedure

Procedures were all identical to those used in Experiment 5.

Processing of data and exclusions

Processing of data took place as in previous experiments (see Section 9.2.4, p. 114).

Training. Exclusions were planned for participants with $< 75\%$ accuracy on the final 2-AFC task, separately for each day. This resulted in the exclusion of six participants. These participants were also eliminated from all subsequent lexical configuration and engagement analyses.

Lexical engagement task. This left 54 participants, whose 3672 (68×54) trials were processed as follows:

- 66 incorrect responses (1.80% of trials) removed;

- 141 trials removed by a $M \pm 3SD$ trim on PL and RT (as previously, per condition);
- 368 trials removed by excluding a further seven subjects, for having less than 70% of their trials remaining in any design cell.

This left 916 *Long ago*, 1086 *Yesterday* and 1092 *Today* word trials. On average, the remaining 47 participants therefore contributed 96.8% of their trials for analysis.

Lexical configuration task. As in Experiment 5, participants excluded from the lexical engagement task were still allowed to contribute their lexical configuration data. However, a further single participant was dropped due to a computer crash and incomplete data. Therefore, this dataset consisted of 53 participants.

13.3 Results

Training was conducted in E-Prime (Schneider et al., 2002). Mouse tracking data were collected in MouseTracker (Freeman & Ambady, 2010). All analyses were performed in R (R Core Team, 2021). Data were visualised with `ggplot` (Wickham, 2016), and the mouse tracking data were processed with `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff et al., 2020).

13.3.1 Training

As in Experiment 5, the analysis that follows is of the 2-AFC task given as the final block of training. On average, across all days and lists, participants paired the correct referent to the heard label 91.4% (SD = 2.80%) of the time. This performance can be seen in Fig. 13.1 (p. 182).

ANOVAs were performed to test for equal list performance on each day[2]. Words learnt on both days showed equivalent performance across lists (words learnt before testing: $F(2, 51) = 1.64$, $p = 0.203$, NS, $\eta_g^2 = 0.061$; words learnt on the day of testing: $F(2, 51) = 1.87$, $p = 0.165$, NS, $\eta_g^2 = 0.068$). Therefore, all further analyses collapsed across training lists.

13.3.2 Lexical engagement

Overview of the mouse tracking measures

The analysis that follows takes the same form as that in Experiment 5. Descriptive statistics for the mouse tracking measures are shown in Table 13.1 and Fig. 13.2 (pp. 182 and 183).

As previously, novel words were first compared across days and competition in a 2 × 2 repeated measures ANOVA for each measure (Table 13.2, p. 184). A second set of ANOVAs then included the familiar words (Table 13.3, p. 185), and post-hoc t-tests were conducted as appropriate.

[2]Note that day could not be entered into the ANOVA as the rotation across lists was not the same for all participants, e.g., participant 1 first learnt words on List 1, then learnt the words on List 2; however, participant 41 first learnt words on List 3, then on List 1

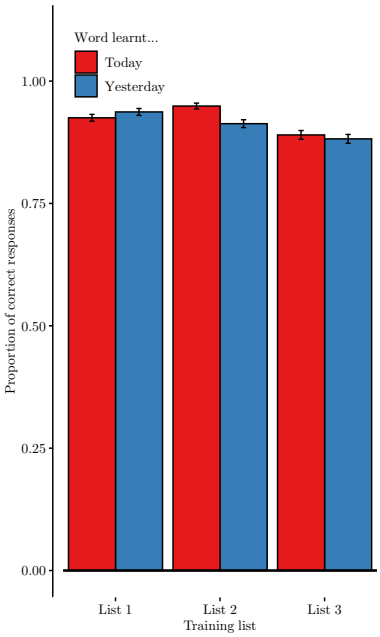
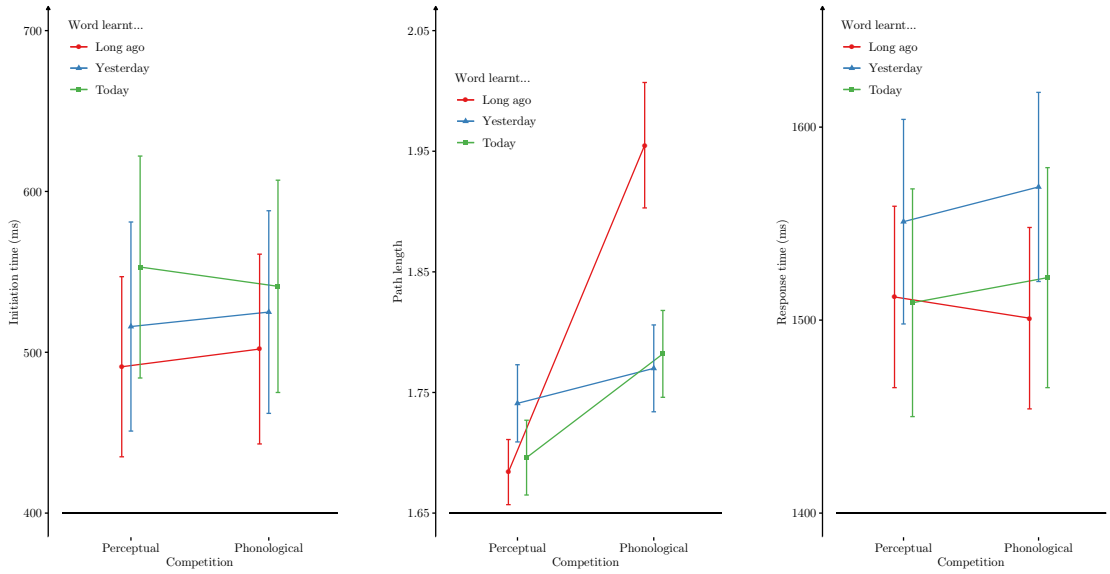


Figure 13.1: Training task (2-AFC) performance in Experiment 6. Error bars show ±1 *SE*

Table 13.1 Summary of descriptive statistics for each lexical engagement measure, by when the word was learnt, and competition, in Experiment 6

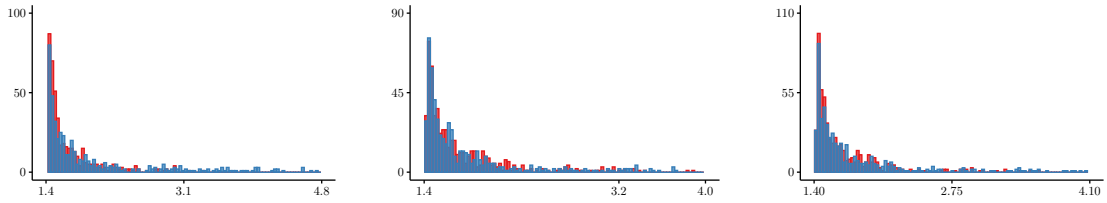
Word learnt...	**Measure**	**Competition**			
		Perceptual		*Phonological*	
		M	*SD*	*M*	*SD*
Long ago	*IT (ms)*	491	383	502	405
	PL	1.68	0.187	1.96	0.354
	RT (ms)	1512	321	1501	325
Yesterday	*IT (ms)*	516	444	525	431
	PL	1.74	0.219	1.77	0.249
	RT (ms)	1551	361	1569	338
Today	*IT (ms)*	553	473	541	453
	PL	1.70	0.213	1.78	0.248
	RT (ms)	1509	401	1522	388

Note. PL units are arbitrary.



(a) IT (b) PL (c) RT

Figure 13.2: Means' plot for each mouse tracking measure in Experiment 6



(a) Familiar (b) Yesterday (c) Today

Figure 13.3: Histograms for PL for each word type in Experiment 6. Perceptual competition trials shown in red; phonological competition trials shown in blue

Distributional analysis.

Spatial responding to all types of words was unimodal, according to Hartigans' dip statistic (all $ps \geq 0.845$, *NS*). This was confirmed by visual inspection of the histograms (see Fig. 13.3, p. 183), which also showed skewed data. However, as in previous experiments, this was not considered to be a problem for the following parametric statistics (Blanca et al., 2017; Ghasemi & Zahediasl, 2012; Lumley et al., 2002).

Initiation time

Descriptive statistics. As in previous experiments, IT showed very small numerical differences (see Table 13.1 and Fig. 13.2, pp. 182 and 183). ITs were approximate to these seen in other experiments without an initiation time cut (Experiments 3 and 4), and much longer than those seem in Experiment 5, with one applied.

Table 13.2 Summary of the day by competition ANOVAs for novel word trials for each of the mouse tracking measures in Experiment 6

Measure	Effect	F	p	η_g^2
IT	*Day*	2.15	0.149, *NS*	< 0.001
	Competition	< 0.007	0.932, *NS*	< 0.001
	Day × *Competition*	0.609	0.439, *NS*	< 0.001
PL	*Day*	0.633	0.430, *NS*	0.001
	Competition	15.1	0.001^{***}	0.015
	Day × *Competition*	2.89	0.096, *NS*	0.004
RT	*Day*	4.80	0.034^{*}	0.004
	Competition	0.872	0.355, *NS*	< 0.001
	Day × *Competition*	0.018	0.894, *NS*	< 0.001

Note. $df = (1, 46)$. Three asterisks (***) denotes significance at the 0.001 level. A single asterisk (*) denotes significance below the 0.05 level.

Yesterday/Today novel word comparisons. The novel word ANOVA showed no main effects, and no interaction (all ps ≥ 0.149). Novel word trials were therefore collapsed across day of learning for comparison with familiar words. The ANOVA is summarised in Table 13.2 (p. 184).

Collapsed novel and familiar word comparisons. The collapsed novel and familiar words were subjected to a 2 × 2 word (familiar, novel) by competition (perceptual, phonological) repeated measures ANOVA (Table 13.3, p. 185). This showed no main effect of competition, or an interaction (both ps ≥ 0.566). This confirmed that competition effects in the other measures were not due to selectively different ITs across conditions. However, a weak main effect of day was observed ($F = 5.00$, $p = 0.030$, $\eta_g^2 = 0.002$). This reflected a trend for participants to initiate movement later for novel words.

Path length

Descriptive statistics. For all types of words, PL was, on average, numerically larger when competition was phonological (see Table 13.1 and Fig. 13.2, pp. 182 and 183). As in Experiment 6, words learnt long ago seemed to produce a stronger competition effect.

Table 13.3 Summary of the day by competition ANOVAs, for words learnt long ago and recently, in Experiment 6, for each of the mouse tracking measures

Measure	Effect	F	p	η_g^2
IT	*Day*	5.00	0.030^{*}	0.002
	Competition	0.334	0.566, *NS*	< 0.001
	Day × *Competition*	0.276	0.602, *NS*	< 0.001
PL	*Day*	26.8	$< 0.001^{***}$	0.021
	Competition	49.9	$< 0.001^{***}$	0.099
	Day × *Competition*	27.3	$< 0.001^{***}$	0.045
RT	*Day*	3.71	0.028^{*}	0.004
	Competition	0.222	0.640, *NS*	< 0.001
	Day × *Competition*	0.312	0.733, *NS*	0.003

Note. df = (1, 46) for IT and PL effects. RT day df = (1, 46); other RT effects df = (2, 92). Three asterisks (***) denotes significance below the 0.001 level; one asterisk (*) denotes significance below the 0.05 level.

Yesterday/Today novel word comparisons. As with IT, PL showed no main effect of day of learning, or a day × competition interaction. Novel words were therefore again collapsed over the two days. However, there was evidence of competition ($F = 15.1$, $p < 0.001$, $\eta_g^2 = 0.015$). The ANOVA is summarised in Table 13.2 (p. 184).

Collapsed novel and familiar word comparisons. The PL data were subjected to the same 2 × 2 word (familiar, novel) by competition (perceptual, phonological) repeated measures ANOVA as IT (Table 13.3, p. 185). This showed main effects of day of word learning, and competition, and an interaction. The presence of an interaction suggested that words learnt long ago produced a larger competition effect, which was consistent with Experiment 5 and Weighall et al. (2017).

Post-hoc *t*-tests. Post-hoc t-tests were performed to further explore the competition effect. This showed that both words learnt long ago ($t = 6.57$, $p < 0.001$, $d = 0.852$) and words learnt recently ($t = 3.89$, $p < 0.001$, $d = 0.262$) engaged in competition, and that the competition effect was stronger for words learnt long ago. Effect sizes were slightly reduced compared to previous experiments.

Response time

Descriptive statistics. Unlike in Experiment 5, which showed evidence of competition (i.e., phonological trials > perceptual trials) for all words, the RT data in Experiment 6 suggested facilitation for words learnt long ago, and competition on novel word trials (see Table 13.1 and Fig. 13.2, pp. 182 and 183).

Yesterday/Today novel word comparisons. The novel word ANOVA, seen in Table 13.2 (p. 184), showed no main effect of competition or an interaction on RT data. However, there was a main effect of day ($F = 4.80$, $p = 0.034$, $\eta_g^2 = 0.004$), and so for the comparison with words learnt long ago, data were not collapsed over day of word learning on the RT measure.

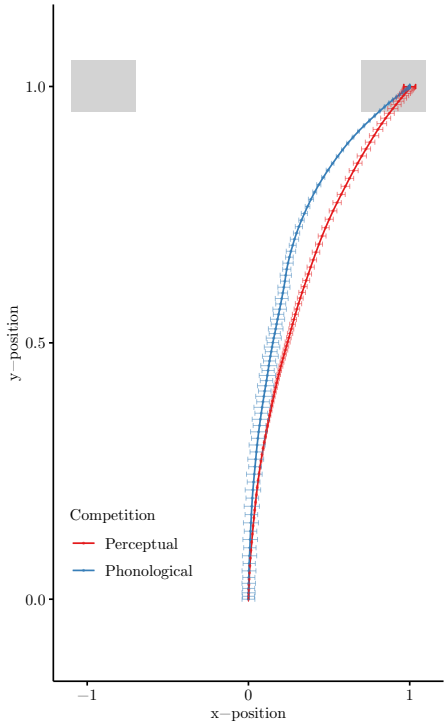
Novel and familiar word comparisons. A 2 × 3 repeated-measures ANOVA was performed (Table 13.3, p. 185), comparing Competition at all three levels of day of word learning. The RT data again showed no main effects of competition or an interaction, but there was a main effect of day ($F = 3.71$, $p = 0.028$, $\eta_g^2 = 0.004$). This was driven by slower response times for novel words, although surprisingly, responses on yesterday word trials were slower than responses on today word trials.

Trajectory analysis

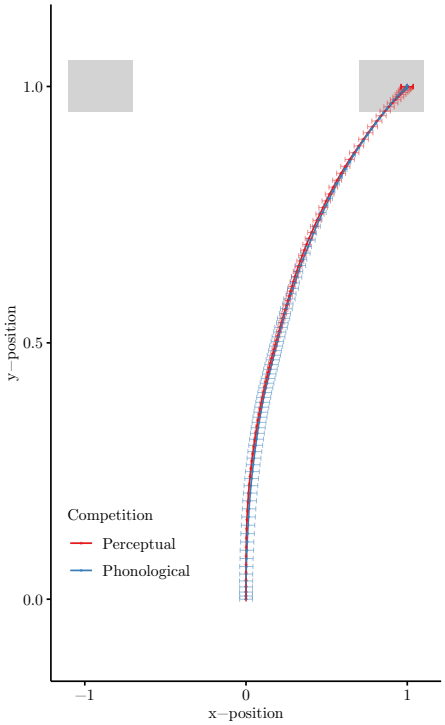
Averaged trial trajectories. Trajectories by when a word was learnt are shown in Fig. 13.4 (p. 187). The pattern of the trajectories was very similar to Experiment 5: there seemed to be displacement towards a phonological competitor for words learnt long ago, and for words learnt today, but not for words learnt yesterday. Similarly, and again as in previous experiments, the displacement was visibly smaller for words learnt today. In terms of their shape, trajectories were more like that seen in Experiments 3 and 4 (Figs. 9.3 and 10.4, pp. 122 and 138), where there was also no IT cut, rather than in Experiment 5 (Fig. 12.4, p. 170).

Standardised time bin analysis. With unimodal data, a comparison of the x-position in phonological and perceptual trials at each of the 101 time bins was performed for all words. Only runs of more than eight bins were accepted as significant, to control for multiple comparisons (Dale et al., 2007). The trajectories are shown in Fig. 13.5 (p. 188).

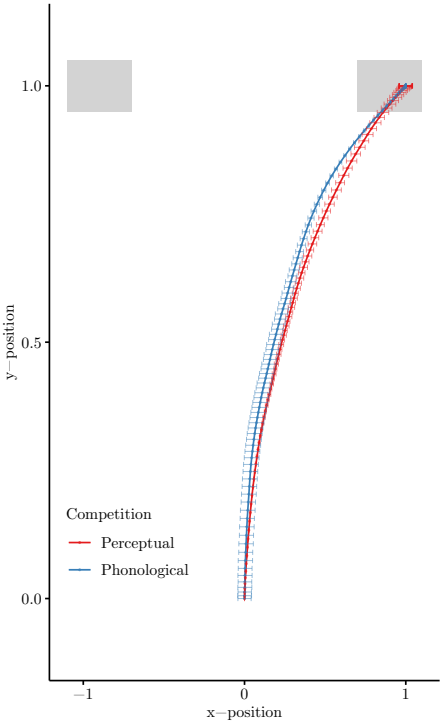
- **Words learnt long ago.** For words learnt long ago, a significant difference in x-position was observed across competition types from the 21st to the 84th time bin (see Fig. 13.5a).
- **Words learnt yesterday.** The t-tests for words learnt yesterday (Fig. 13.5b) showed no significant differences in position across competition trial types. This mirrors the pattern observed in Experiment 5.
- **Words learnt today.** There was a long run of significantly different x-positions, between the 12th and the 80th time bin (Fig. 13.5c).



(a) Words learnt long ago

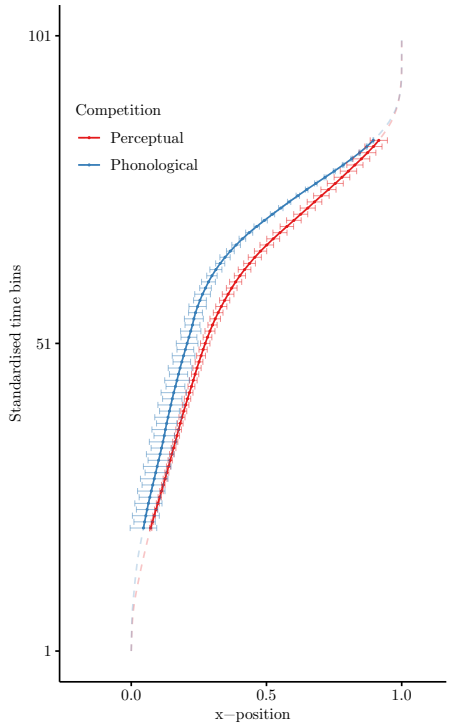


(b) Words learnt yesterday

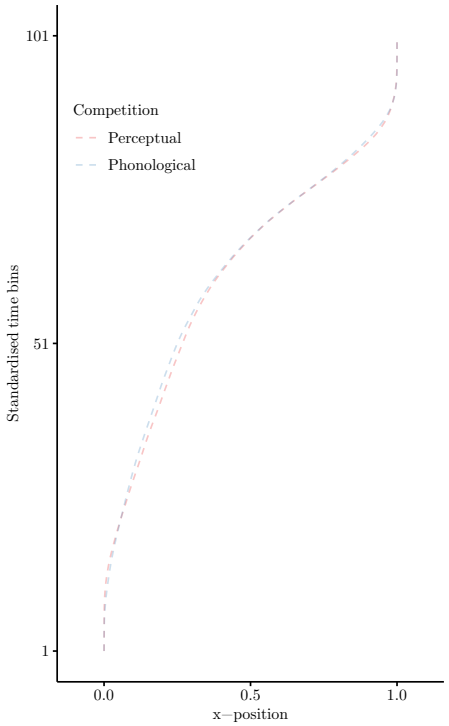


(c) Words learnt today

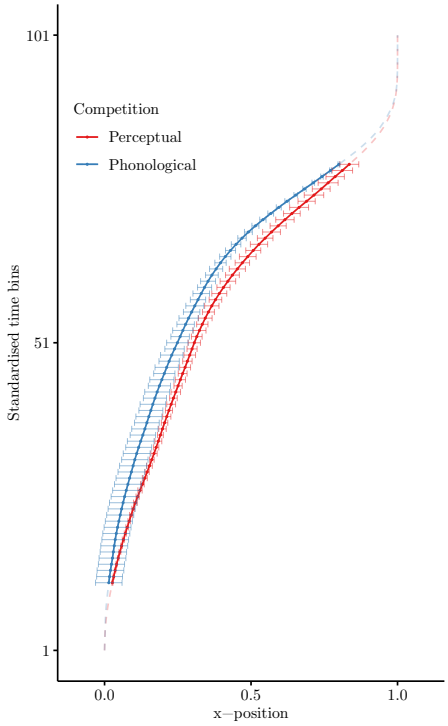
Figure 13.4: Averaged participant trajectories plotting x- against y-position for each time of word learning in Experiment 6. Each point represents a single time bin. Error bars represent $\pm$ 1SE in the x-position



(a) Words learnt long ago



(b) Words learnt yesterday



(c) Words learnt today

Figure 13.5: Averaged x-position against standardised time in Experiment 6, for words learnt at each time point. The solid saturated line (if present) indicates significantly different positions across conditions in each time bin (paired-samples t-test; $p < 0.05$). Each point represents a single time bin. Error bars represent $\pm$ $1SE$ in the x-position

13.3.3 Lexical configuration

Analyses of the lexical configuration data were as in Experiment 5, comparing the two recall tasks (stem completion and picture naming), for words learnt either yesterday, or today.

Descriptive statistics

Performance was best for words learnt yesterday, on the stem completion task ($M = 49.6\%$, $SD = 22.7\%$). However, performance for words learnt today on the stem completion task ($M = 31.3\%$, $SD = 18.0\%$) was still better than for either words learnt yesterday ($M = 19.8\%$, $SD = 17.8\%$) or words learnt today ($M = 21.5\%$, $SD = 19.4\%$) on the picture naming task.

Inferential statistics

A 2 × 2 repeated measures ANOVA was conducted, inputting the two tasks and each type of novel word. As in Experiment 5, this showed main effects for when the words were learnt ($F = 16.8$, $p < 0.001$, $\eta_g^2 = 0.044$) and the task testing learning ($F = 118$, $p < 0.001$, $\eta_g^2 = 0.207$), as well as an interaction between the two ($F = 88.1$, $p < 0.001$, $\eta_g^2 = 0.062$).

Post-hoc t-tests again showed an identical pattern to that observed in Experiment 5. Performance was flat across days for picture naming ($t = -0.828$, $p < 0.411$, *NS*, $d = 0.088$), but not for stem completion ($t = 7.15$, $p < 0.001$, $d = 0.876$). Stem completion performance was superior to picture naming performance for words learnt today ($t = 5.45$, $p < 0.001$, $d = 0.523$) and yesterday ($t = 12.5$, $p < 0.001$, $d = 1.42$).

13.4 Discussion

Experiment 6 tested whether, the competition effects driven by newly learnt words were semantic (and therefore lexical). Experiment 6 showed a pattern of competition effects similar to that observed in Experiment 5: competition effects were again observed for the PL measure for both novel words and familiar words, with stronger effects for familiar words. Participants showed greater attraction to the on-screen competitor on phonological competition trials compared to perceptual competition trials. The key difference between Experiment 5 and 6 is that both the phonological and perceptual competition conditions for novel words in Experiment 6 used a base word target that now had a newly trained phonological competitor. That a competition effect was still observed in Experiment 6 suggests that this effect is not simply due to the target base word evoking a recently learnt novel form in the phonological competition condition alone: in Experiment 6 this was able to happen in both phonological and perceptual competition conditions. Instead, the PL data suggest that participants processed the relationship between the evoked novel form and the on-screen competitor. Where the novel competitor object was the referent of the evoked novel word form, as in the phonological competitor condition, this resulted in greater competition. This suggests that the participants represented

semantic information during word learning, and that this information was accessible in the lexical engagement task. Notwithstanding the fact that familiarity was a confounding variable, the remarkable similarity in the broad pattern of results between Experiments 5 and 6 invites one to tentatively conclude that the novel word representations demonstrating competition effects *are* lexical in nature.

There were however some discrepancies in the findings between Experiments 5 and 6. First, in the present experiment there was no competition effect in the RT data. This may be as RT is a less sensitive and reliable measure (relative to PL) for detecting the effects of interest in mouse tracking studies (cf., Maldonado et al., 2019), as noted in Chapter 12 as well. Secondly, there was also a main effect of IT, although recall that in the present study there was no IT cut. One possibility is that these factors are related: without an IT cut participants were not under pressure to immediately start moving the mouse. They therefore may have paused for slightly longer (though at most only $\sim$ 60ms) on novel word trials compared to familiar word trials because on the former there was a novel object on screen, whereas on the latter all on-screen objects were familiar. Participants may have needed slightly longer to process the novel objects. Another discrepancy is the divergence between the spatial and temporal data – participants moved much further on phonological trials, but seemed to do so in the same amount of time. This occurred despite otherwise very typical trajectories, much like that seen in previous experiments. However, there shapes of the trajectories were a little difference, though this again seems likely due to the missing IT cut, as the trajectories in Experiment 6 were quite similar to the trajectories in Experiment 4, and the second mode of Experiment 3.

Finally, the pattern of the lexical configuration data was identical to that seen in Experiment 5. This is unsurprising, given the training and lexical configuration tasks were also identical. However, it is interesting that the task which is more 'semantic' – picture naming – so consistently has shown little improvement across days, and yet recall of the form (when cued by the stem) showed such dramatically large improvements with a night of sleep. Also interesting is how this task dissociates from the apparent pattern in the lexical engagement task – where a night of sleep seems to make little difference in both experiments, as yesterday and today word trials were statistically indistinguishable. The issue of how semantic information is handled by the cognitive system is addressed further in Experiment 7, which also sought to address a limitation of Weighall et al. (2017) discussed in previous chapters (Chapters 11 and 12, pp. 149 and 155), and the confounding familiarity variable present in the current experiment.

CHAPTER

FOURTEEN

EXPERIMENT 7
THE NATURE OF NOVEL WORD REPRESENTATIONS: THE LEXICALITY OF NOVEL WORDS

14.1 Introduction and rationale

Experiment 7 was the final of three mouse tracking experiments run to explore lexical engagement by means of lexical competition immediately after word learning.

Experiment 5 had replicated a lexical competition effect in the literature (Weighall et al., 2017) and adapted this effect to mouse tracking. However, given the specific comparisons made by Weighall et al., and the conditions used in Experiment 5, it is debatable whether the competition that was observed was *lexical* in nature. In Experiment 5, the phonological and perceptual competition conditions differed according to whether the target base word had a newly trained phonological competitor, or not. It is possible that the competition effects emerged due to a stem shared by both a learnt novel competitor and the familiar target, with no reference to the on-screen objects, and therefore to the novel word's semantics.

Experiment 6 allowed this effect to be investigated further, by testing if participants were sensitive to semantics in the lexical engagement task. In that experiment both the phonological and perceptual competition conditions used a target base word that had a newly trained phonological competitor: the conditions differed according to the on-screen competitor objects. Although confounded by the familiarity of those objects, the results of that experiment at least suggested that participants were sensitive to the semantics of the novel words. Relative to the phonological competition condition, less attraction to the super-novel competitor objects was observed in the perceptual competition condition, which suggested that participants were not mapping the novel word evoked by the shared stem to them. However, this may simply have been because participants realised that they had not seen the super-novel object before, and therefore inferred that it could not be the referent for the novel label evoked by the target base word. Such a process, based on the lack of familiarity with the super-novel objects, would not have required participants to have bound the newly learnt novel label to its referent in a single lexical representation.

Experiment 7 was the final experiment of the set, and addresses the above confound in Experiment 6. In Experiment 7, in both the phonological and perceptual competition conditions, the target base word had a newly trained phonological competitor (whose referent did not appear on perceptual competition trials). Further, in both conditions the on-screen competitor object was a novel object that participants had learnt a label for during training. However, whereas in the phonological competition condition the novel word evoked by the base word (e.g., 'aliet', when hearing "Click on the alien") was the label for the on-screen novel object (i.e., ALIET), in the perceptual competition condition it was not (e.g., 'angesh' evoked by "Click on the angel", but BADMINTEEF present as the competitor object, with 'badminteef' also having been learnt). This design allowed one to test the lexicality of the representations underlying the competition effects observed in Experiments 5 and 6.

If a competition effect was observed in Experiment 7, this would lead directly to the conclusion that, at the very least, the stem of a novel word was correctly bound to the referent object. This would imply that the representations engaging in lexical competition were 'lexical'[1] immediately after learning. This is a conclusion that thus far is not clear in the literature, although there have been some suggestions that it is the case, with various 'word-like' properties being demonstrated (Bartolotti & Marian, 2012; Lindsay & Gaskell, 2013; Kapnoula & McMurray, 2016a; McMurray et al., 2017; Tham et al., 2015; Weighall et al., 2017). Alternatively, if no competition effect was observed, then this would imply that participants were not binding the stem of the newly learnt novel words with the correct referent object. This in turn would suggest that the competition effects observed in Experiments 5 and 6 were not the result of lexical representations, in which form and meaning were bound into a single representation. Instead, this would suggest those competition effects were the result of non-lexical processes.

Testing for Experiment 7 began in March 2020. However, the University of Hull suspended research activity within a few days of testing beginning, due to the CoViD-19 pandemic. From then until the thesis submission date (23rd April 2021), further research activity was not possible, and therefore, only a single participant was tested (in March 2020). Further testing to complete the experiment is planned (commencing October 2021).

Chapter 15 is the next and final chapter of this thesis, and will summarise the findings and consider further the nature of novel word representations. The methods for Experiment 7 are outlined below.

14.2 Methods

14.2.1 Participants

It is anticipated that Experiment 7 will use a recruitment strategy and sample size as outlined in Experiment 5. Participants are planned to be excluded during testing if they are to be inaccurate responders, and replaced until a target sample size of 60 participants is achieved. Participants will not have participated in any previous

[1]Insofar as they contained a form bound to a referent, and insofar as words are a bundling of referent and form

experiments, or have any confounding disorders (e.g., sensory, learning or language difficulties). They will ordinarily use the mouse with their right hand.

Testing will be in accordance with procedures approved by the Faculty of Health Sciences ethics committee at the University of Hull. They will volunteer their time in exchange for course credits.

14.2.2 Materials and apparatus

Materials will be exactly as in Experiments 5, and Weighall et al. (2017). For a full list, see Tables D.1 to D.3 and D.5 (pp. 219–222).

14.2.3 Design

Unlike Experiment 6, but like Experiment 5, an IT cut will be present in Experiment 7, in line with best practice Kieslich, Schoemann et al. (2020). This will mean that if participants do not initiate movement before 450ms elapse, they will see a warning on-screen, and that trial will be discarded.

The structure of the experiment will also mirror Experiments 5 and 6. It will run over two days, with tasks in the following order: training on the first day, and then on the second day, further training, a lexical engagement task, and two lexical configuration tasks.

The only difference in Experiment 7 will be the novel word perceptual competition condition. Consider a future participant tested on list A123. She will learn List 1 words on the first day of training, and List 2 words on the second day, when testing will also take place. In the new design, List 3 words are not used at all, for this participant. Instead, for words learnt yesterday perceptual competition trials, the participant will see a List 1 base word object (e.g., ALIEN) against a List 1 novel competitor with a non-overlapping label (e.g., NAPKIG). As before, the participant will have to respond to the base word, and the will hear the form 'alien'. This will contrast from Experiment 5, where participants saw a List 3 base word object (e.g., BALCONY) against the List 1 novel competitor, and from Experiment 6, where the List 1 base word was placed against an un-named super-novel referent.

14.2.4 Procedure

The procedure will be identical to that used in Experiments 5 and 6. Training will be conducted in E-Prime (Schneider et al., 2002). Mouse tracking data will be collected in MouseTracker (Freeman & Ambady, 2010). All analyses will be performed in R (R Core Team, 2021). Data will be visualised with `ggplot` (Wickham, 2016), and the mouse tracking data will be processed with `mousetrap` (Kieslich & Henninger, 2017; Kieslich, Wulff et al., 2020).

Processing of data and exclusions

Processing of data will take place as in previous experiments (see Section 9.2.4, p. 114). Additionally, trials will be filtered out according to the IT cut.

Training. Exclusions are planned for participants with $< 75\%$ accuracy on the final 2-AFC task, separately for each day.

Lexical engagement task. Trials will be excluded for:

- Incorrect responses;
- Exceeding the IT cut of 450ms;
- Exceeding a $M \pm 3SD$ trim on path length and response time measures (as previously, per condition);

Following these procedures, participants will be examined for how many trials they have remaining. Any participants with $< 75\%$ of their trials left, in any condition, will be removed. Participants removed from the lexical engagement task will be allowed to provide lexical configuration data, given sufficient 2-AFC performance during training.

CHAPTER

FIFTEEN

GENERAL DISCUSSION

Human word learning was the topic of this thesis. Word learning is important as words are the fundamental building blocks of language, and language is interesting as a uniquely human capacity (Hockett, 1960; Pinker, 1995; Rivas, 2005). Historically, there has been considerable debate about how language is acquired, and how words are learnt (Chomsky, 1959; Skinner, 1957). To become 'word-like', a novel utterance must undergo cognitive processing, the mechanisms of which are still debated (e.g., Davis & Gaskell, 2009; Goldinger, 1998; Kapnoula & Samuel, 2019; Lindsay & Gaskell, 2010; McClelland et al., 1995; McMurray et al., 2017; Palma & Titone, 2020).

15.1 Summary of thesis findings

15.1.1 Experiments 1 and 2

The first experiments investigated the 'fast mapping' (FM) phenomenon, which has become a recent subject of interest in the adult word learning literature (e.g., Cooper et al., 2019a, 2019b, 2019c; Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017; Merhav et al., 2014, 2015; O'Connor & Riggs, 2019; Sharon et al., 2011). FM is an experimental procedure in the developmental literature used to simulate the word learning environment of early childhood (Carey & Bartlett, 1978; Gernsbacher & Morson, 2019). In a typical setup, a novel object is placed against a familiar object, and an experimenter pronounces a novel word – in, for example, the context of a request for the child to hand over an object (e.g., Dysart et al., 2016; Riggs et al., 2015). Crucially, the child has not heard the novel word or seen the novel object before, but without prompting, maps the novel word to the novel object, and passes it to the experimenter. This behaviour is termed 'referent selection'. Next, possibly after a retention interval, the child is asked the same question in the presence of a series of other novel objects, which again were not previously seen, and an object from the referent selection trial. This allows the experimenter to test the child's retention of the novel word and the robustness of the 'fast mapped' word-referent link.

FM is of interest to researchers of adult word learning as it gave rise to findings

that were inconsistent with a complementary learning systems account. The complementary learning systems model (CLSM; McClelland et al., 1995), a prominent model of memory, had recently been applied to word learning (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). This account models word learning in two stages: episodic representations are first put into their own store, and only later incorporated into long term memory. This 'siloing' of episodic representations allows details relevant to future word use to be extracted and generalised across the instances in which it was encountered. For example, encountering the word 'cat' more often in the presence of the domestic house cat than in the presence of a violin may cause the semantic representation CAT and phonological representation /kæt/ to become paired in the mind of a learner (e.g., Hawkins & Rastle, 2016). Moreover, details which were not relevant to productive use of the word, such as whether the word was whispered or shouted, need not be maintained in the long term (though see Pufahl & Samuel, 2014). Secondly, the CLSM also provides a solution to the stability/plasticity dilemma (Carpenter & Grossberg, 1988), which describes the contrast in a cognitive system between needing to add new information easily, but also store old information in stable networks. It is postulated that information siloed in an episodic store, thought to centre on the hippocampus (e.g., Gabrieli et al., 1988; Scoville & Milner, 1957), is then consolidated into an abstract, generalised, cortical store through the process of *reinstatement*. It has been suggested that sleep is one such time when reinstatement may occur (Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010), resulting in the consolidation of new memory traces.

The FM findings broke with this model in two ways. Firstly, patients with brain injuries (specifically, with hippocampal damage) were found to selectively learn better through FM (e.g., Sharon et al., 2011; Merhav et al., 2015). This suggested that FM allowed some bypassing of the episodic silo, which the original formulation of CLSM required (McClelland et al., 1995). However, other research groups failed to replicate the finding in similar patients (Warren & Duff, 2014; Warren et al., 2016), older adults with naturally and similarly reduced hippocampal volumes (Greve et al., 2014), and in other patient groups (Korenic et al., 2016; Sakhon et al., 2018).

Secondly, data were presented suggesting that sleep was not a prerequisite for consolidation, as evidenced by early 'lexical engagement' – suggesting that the process described by the original formulation of CLSM was inaccurate, at least under particular learning conditions (Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017; Zaiser et al., 2019b). Lexical engagement is the ability of a word representation to interact with other word representations (e.g., Kapnoula et al., 2015; Leach & Samuel, 2007). By comparing the interactions observed between known words, and the interactions observed between a known word and a learnt word, one has an implicit measure of how 'word-like' a newly-learnt word is. Lexical engagement may take many forms (each relating to a different aspect of a lexical representation; McMurray et al., 2017). Coutanche and Thompson-Schill (2014) used a lexical competition paradigm in the visual modality, showing where a novel and competing orthographic form had been trained (e.g., 'torato', for 'tomato'), responses to the familiar word target were slowed. No such slowing occurred for words which had not had a competitor trained. Previous work had demonstrated that a lexical competition effect only emerged after sleep and further training (Bowers et

al., 2005), and furthermore, that the effect was driven by lexical processing (Dumay & Gaskell, 2012; Qiao et al., 2009). Competition was caused by the word representations being linked and co-activated, making word recognition processes more inefficient. However, Coutanche and Thompson-Schill found that for words trained by FM, lexical engagement was present immediately after training. This was consistent with recent updates to the CLSM (e.g., McClelland, 2013; McClelland et al., 2020), and with work on schema in rats (Tse et al., 2007), which showed that where information was schema consistent, it could be integrated more rapidly and thus competition effects would emerge earlier. This argument was articulated to also explain the patient data (Atir-Sharon et al., 2015; Merhav et al., 2014, 2015; Sharon et al., 2011). Further work showed that the competition effect was related to how much weight individual participants gave to semantic information, and how strongly the schema was activated (Coutanche & Koch, 2017). Other authors have also found that schema are supportive of word learning (Havas et al., 2018).

It was against this background that Experiments 1 (Chapter 5, p. 55) and 2 (Chapter 6, p. 65) were carried out. Experiment 1 was planned to extend the Coutanche and Thompson-Schill (2014) effect by determining whether a specific reference to a feature common to the two on-screen referents (intended to activate the schema) was required. For example, when learning a novel insect, Coutanche and Thompson-Schill used a question of the form 'Are the antennae of the 'torato' pointing down?', with TORATO and GRASSHOPPER on screen. Another experiment in their paper had suggested that when the grasshopper was removed, the competition effect was not observed before sleep. This suggested that the reference to schema-relevant features within the question was irrelevant, and that the competitor object activated the schema for the integration of novel material. Experiment 1 tested this further by asking a question which required participants only to disambiguate the referent: the question was 'Where is the 'fostil'?', the answer to which was 'On the left/right of the screen' (see Fig. A.1, p. 213). However, under this training regime, Experiment 1 failed to find evidence for Coutanche and Thompson-Schill's key finding of pre-sleep lexical competition.

Having failed to extend their work, a replication of the study was run. Being closer to an exact methodological replication, it was anticipated that a failure to replicate in this second experiment would align with the more recent literature suggesting that the reported FM effects are non-replicable[1]. However, a successful replication would have indicated that the FM effects were only robust under particular training conditions. However, Experiment 2 also showed no evidence of pre-sleep lexical engagement.

15.1.2 Experiments 3 and 4

Following the failures to extend and replicate the FM effects in Experiments 1 and 2, the research project that had been planned had to be changed at short notice, as it no longer made sense to run studies following up such a fragile (if at all real) effect. However, this was only true as it related to FM – evidence for pre-sleep lexical

[1] Although as of September 2021, no replication of the lexical *engagement* effects has been published, though cf., Cooper et al. (2019a)

engagement had been steadily building in the literature for the previous decade (see Table 7.1, p. 85, McMurray et al., 2017; Palma & Titone, 2020). However, only half of these papers focussed on lexical competition. The FM literature in support of designing the first two experiments also emphasised the importance of semantics, but there were contradictory findings in the rest of the word learning literature about the importance of semantic information during learning (Dumay et al., 2004; Hawkins et al., 2014; Hawkins & Rastle, 2016; Henderson et al., 2013). A plan of research was therefore made continuing the themes of the FM work, investigating the nature of novel word representations under conditions in which semantics were also learnt.

Through the review of the pre-sleep lexical engagement literature, it was noted that many of the more recent papers used eye tracking (Bartolotti & Marian, 2012; Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Weighall et al., 2017). To explain the finding of early lexical engagement, when other authors had not found it (e.g., Bowers et al., 2005; Gaskell & Dumay, 2003; Dumay et al., 2004; Dumay & Gaskell, 2007), researchers reporting these pre-sleep effects suggested that their findings might result from the more specific activation of competitors that a paradigm like eye tracking allows (e.g., Kapnoula et al., 2015; Kapnoula & McMurray, 2016a; Kapnoula & Samuel, 2019; Weighall et al., 2017). Moreover, it was argued that, whereas previous measures indexed the *overall* level of lexical activity, eye tracking tracked specifically the rising and falling of any on-screen referent's activation over time. It was noted that similar arguments were made for mouse tracking, with the additional benefit that mouse tracking was a truly continuous and graded measure (Bartolotti & Marian, 2012; Spivey et al., 2005).

However, in switching to a novel paradigm, it had to be piloted. To do so, two experiments were conducted, looking for a competition effect between known words. Various aspects of the design, such as how best to organise trials and which measures to take, were unclear, and highly variable (see Chapter 8, p. 93). Experiment 3 (Chapter 9, p. 107) demonstrated that mouse tracking was a viable proposition and suggested the best measures to use in further mouse tracking work. Experiment 4 (Chapter 10, p. 129) confirmed these suggestions, and further showed that a design and some stimuli used in the literature for studying novel words (Weighall et al., 2017) were amenable to mouse tracking.

15.1.3 Experiments 5, 6 and 7

Experiments 5–7 applied the mouse tracking protocols that had been designed in Experiments 3 and 4 to novel word learning. Implementing the eye tracking design of Weighall et al. (2017), the experiments looked for an immediate lexical competition effect, and asked to what extent this effect was driven by truly word-like representations. Instead of being driven by word-like representations, made up of a referent bound with a label, the alternative possibility was that the words were based on some sort of episodic cueing – with participants recalling that they had recently learnt a similar sounding word, and recognising a learnt on-screen referent.

Experiment 5 (Chapter 12, p. 155) was a direct replication of Weighall et al.

(2017), adapted to mouse tracking. The experiment found evidence of immediate lexical competition. With respect to evidence for consolidation (as per the CLSM; McClelland et al., 1995), when participants were cued by the stem of the novel word, their rate of recall improved with sleep. However, this was not the case when they were cued by the referent. The lexical engagement data also showed no evidence of consolidation: there was no statistically significant difference on the most sensitive mouse tracking measure, path length (PL), across the two days. This pattern of lexical configuration and engagement data was entirely consistent with Weighall et al. (2017).

Experiment 6 (Chapter 13, p. 177) suggested that the representations engaging in immediate lexical competition were semantic, as with the same set up as Experiment 5, participants still showed a difference in response profile across conditions where the on-screen referent was either trained, or super-novel. In both cases, a novel competitor was trained for the familiar target object – so if the competitor was evoked by this object's label alone, participants would have mapped it to the novel referents in both conditions – thus showing no effect. The lexical engagement effect was therefore mediated by the semantics of the learnt referent. This had not been shown in previous studies, and was for the first time, a suggestion that the competition observed was indeed lexical, and not based on episodic cueing. However, the design was confounded by the fact that the participants were familiarised with a referent in one of the conditions, but not in the other, and this familiarity may have biased responding. Differences were again not evident across days in the lexical engagement data. Lexical configuration data were as in Experiment 5, with consolidation evident when participants were cued by the stem, but not when cued by the referent. This is again consistent with Weighall et al. (2017).

Experiment 7 (Chapter 14, p. 191) is yet to run and no conclusions can be reached from the single participant from whom data could be collected before the CoViD-19 pandemic forced the suspension of research in March 2020. However, this experiment, when it runs, will give a definitive conclusion as to whether participants are representing semantics in their novel word representations. This in turn will provide evidence for the lexicality of the novel word representations and the competition exhibited.

15.2 Thesis findings in context

15.2.1 Pre-sleep lexical competition – not so surprising?

Experiments 5 and 6 found evidence consistent with pre-sleep lexical competition. A wide body of research now provides consistent evidence for pre-sleep lexical competition effects (see Chapter 7, p. 83), contrary to the predictions of the CLSM. Even beyond such studies, there is additional data that cannot be accounted for by a complementary learning systems account (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). For example, in Bowers et al. (2005), a paper which showed effects related to the sleep-based consolidation of orthographic forms, participants were tested over three testing sessions: on the first day after training, on a subsequent day before further training (but after sleep) and on the same subsequent day, after

training. Testing used a response time (RT) task: participants had to categorise targets as artefacts or natural objects, and targets had either had a competitor trained, or had not. The first session showed no competition effect (i.e., categorisation RTs were unaffected by whether a competitor had been learnt), but the second and third testing sessions did: consistent with the integration of a competitor word with the target, as predicted by the CLSM. By subject and by item analyses were performed. However, even in the first session (before sleep), there was a 17ms difference between conditions, that was trending towards significance by subjects and items ($p_{subjects} = 0.16$, $p_{items} = 0.06$). Moreover, the *size* of the competition effect only increased significantly between testing sessions one (i.e., before sleep) and three (i.e., after sleep *and* further training) – despite the emergence of statistical significance in the second session (i.e., after sleep but *before* further training). The size of the competition effect in the second session was 33ms, growing to 48ms in the third session. This does not suggest that a qualitative change happened with sleep (as suggested by the binary categorisation of 'competition present' or 'competition absent'), but rather that there was an emerging trend towards lexicality that began on the first day. Whilst this is consistent with later research showing pre-sleep competition effects (e.g., McMurray et al., 2017; Palma & Titone, 2020), it is inconsistent with accounts such as that proposed by Davis and Gaskell (2009), Dumay and Gaskell (2007) and Lindsay and Gaskell (2010). Additionally, it is notable that at all stages, training seemed to produce an equal increase in the size of the competition effect[2]. This suggests that further training, *not* sleep, was important. This finding, and the pre-sleep lexical engagement literature, conclusively show sleep to be a 'not necessary' condition of lexicalisation. However, in many papers also fail to show a single night of sleep resulting in behavioural change (cf., Dumay & Gaskell, 2007), implying that sleep is also a 'not sufficient' condition of lexicalisation (Brown et al., 2012; Coutanche & Thompson-Schill, 2014; Gaskell & Dumay, 2003; Hawkins & Rastle, 2016; Henderson et al., 2013, 2014; Himmer et al., 2017; Walker et al., 2019).

Building on this, it should also be emphasised that the finding of word-like behaviour in very newly learnt words is rather old, and seemingly not accounted for by complementary learning systems accounts (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). Magnuson et al. (2003) showed with eye tracking that newly learnt words interacted with each other similarly to natural words, and their representations also contained information that natural words are sensitive to (e.g., frequency of occurrence; and neighbourhood density effects; cf., Luce & Pisoni, 1998). Unfortunately, the nature of the stimuli meant that competition between these learnt forms and true words could not be tested.

Lastly, whilst pre-sleep lexical *competition* has only been demonstrated relatively recently[3], lexical engagement by other measures has been quite robustly shown over many paradigms and at many levels (e.g., at the morphological level, Lindsay et al., 2012; at the sub-lexical level, Leach & Samuel, 2007; Snoeren et al., 2009; at the semantic level, Geukes et al., 2015; Tham et al., 2015; see Chapter 7, p. 83, McMurray et al., 2017). Whilst one can debate the *nature* of the competition effects

[2]Session one vs. two: 16ms; session two vs. three: 15ms

[3]To the knowledge of the author, first shown by Fernandes et al. (2009).

that have been reported – a point that cannot be addressed by this thesis, given the fact that Experiment 7 has yet to run – lexical engagement has been demonstrated pre-sleep so widely that the evidence for it is now quite conclusive.

15.2.2 Addressing the absence of competition in fast mapping

Given then that pre-sleep lexical competition is not surprising, the failures to find no effect in Experiment 1 and 2 under FM conditions cannot be attributed to consolidation not having taken place. Alternative explanations must be considered. One suggestion in the literature is related to the number of exposures during training. This is discussed below.

In discussing the FM findings, a clear distinction must be drawn between the overlapping claims. Some authors have asserted that FM leads to *better* learning (e.g., Sharon et al., 2011; Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017) – given recent reviews of the work and the failures to replicate, this does not seem to be the case (Cooper et al., 2019a, 2019b, 2019c). 'Better' in this context means either an accelerated emergence of lexical competition – in healthy adults (e.g., Coutanche & Thompson-Schill, 2014; Coutanche & Koch, 2017) – or learning in groups otherwise incapable of efficient learning, such as patients (e.g., Sharon et al., 2011).

However, even if learning by FM is not better, this does not necessarily mean that learning is poor – at least on lexical engagement measures (though there does seem to be agreement in the literature that lexical configuration is poor under FM conditions; compare Coutanche & Thompson-Schill, 2014; Warren et al., 2016). A confounding factor in the FM work is that very few exposures are used in training, and thus the broader question of whether FM training is capable of leading to pre-sleep lexical competition generally is not addressed. Two aspects characterise FM: a small number of exposures, and an inferential mapping between the novel words and referent (cf., Carey & Bartlett, 1978). It is the second of these that may still bring about an effect. While a small number of exposures may not lead to pre-sleep lexical competition, it is still possible that a larger number of exposures combined with an inferential mapping does. Relative to the 12 exposures per novel word used by Weighall et al. (2017), Coutanche and Thompson-Schill (2014) used only two. The findings of Bowers et al. (2005) suggest that training is important, and this is supported by evidence elsewhere in the literature. Lindsay and Gaskell (2013) found evidence of pre-sleep lexical competition where Bowers et al. (2005) had not, with a similar RT paradigm. However, this was only after further rounds of training/testing sessions (10 exposures in a block of training). It may also be that sleep interacts with training – Walker et al. (2019) found that with only five exposures, no sleep-based consolidation took place. Whilst sleep-based consolidation was found for 10 and 20 exposures, there was no significant difference between the two – suggesting some kind of threshold beyond which further training is not helpful. This work used the same semantic categorisation task used by Bowers et al. (2005) and Coutanche and Thompson-Schill (2014).

With 10 exposures, Wang et al. (2017) showed that a competition effect only emerged after sleep. This paper again looked for competition in orthographic word

forms, and used the semantic categorisation task from Bowers et al. (2005) and subsequent researchers. Using a design where participants were tested at 8AM and 8PM, and then at 8AM the following day (AM group), or at 8PM, and then at 8AM and 8PM the following day (PM group), Wang et al. were able to manipulate when sleep occurred in the training/testing cycle and thus assess its contribution independently of training and further time (following Dumay & Gaskell, 2007). Whilst training occurred in the first session, no further session had additional training. Both groups showed competition only in the session following sleep (at 8AM on the following day for both groups; after 12h for the PM group, but 24h for the AM group). However, whilst sleep may help *instead of* further training, this experiment does not preclude the possibility that the system is flexible enough to use *either* further training *or* sleep. Indeed, it may simply be that the consolidation processes suggested to take place during sleep allow for the reactivation of the novel word, much as further training also does.

In summary, whilst there is little evidence for FM promoting better word learning, it may be premature to conclude that it does not. It still remains a possibility that FM learning conditions give rise to better word learning compared to learning conditions without the presence of a competitor (to activate the relevant schema) if the number of training trials is increased.

15.3 Towards a new theory of word learning

15.3.1 'Echoes of echoes': an episodic lexicon

In previous chapters, word learning has been discussed in the context of the CLSM (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010; McClelland et al., 1995), as this account has support in the literature (see Chapter 3, p. 21). However, an alternative account may be a better fit for the data presented in this thesis: an episodic account of lexical access and storage (Goldinger, 1998). Other authors have also tentatively suggested that their data fit this account (e.g., Kapnoula & Samuel, 2019).

The conflict between these two theories concerns abstraction, and the point at which it takes places. Whilst theories of an abstract lexicon predict that abstraction takes place as part of storage, episodic accounts propose that all encounters with a word are stored, and abstraction takes place at recognition. Goldinger (1998) presented such an episodic account after finding that a computational model which implemented such storage correctly predicted his participant data.

The computational model discussed by Goldinger (1998), is purely episodic, and represents one extreme end of the possible episodic/abstractionist spectrum. The model (MINERVA 2; Hintzman, 1986, 1988) assumes that *all* encounters with a word ('episodes') are stored, complete with even those details irrelevant for its productive use (e.g., if spoken, indexical properties such as speaker identity; Kapnoula & Samuel, 2019; or even environmental noises heard at the time of encoding, such as a phone ringing; Pufahl & Samuel, 2014). This results in a large store, including many traces which are redundant (due to the high degree of similarity). Upon perceiving input (e.g., a heard word), the system 'probes' each stored trace, in parallel.

Each stored trace then echoes a reply to the probe, with the echo described by two parameters. The first parameter is the intensity of the echo – that is to say, the degree to which a trace matches the probe. The second is the content of the echo – traces are echoed back in their entirety, including those irrelevant details which nevertheless seem to be stored from perception (e.g., Kapnoula & Samuel, 2019; Pufahl & Samuel, 2014). Abstraction – for example, sufficient to allow the recognition of a word despite it occurring in an unfamiliar voice – therefore occurs as part of word recognition, when the system analyses the echoes. Having received many echoes back, the system averages across them forming a 'generic echo' in working (but crucially, not long term) memory. It is this 'generic echo' that allows for recognition under novel conditions (e.g., a new speaker). This is quite different from abstractionist theories – which predict abstraction occurring prior to retrieval, as part of storage (e.g., McClelland et al., 1995). In the case of the CLSM, McClelland et al. (1995) model abstraction to occur as old and new information is interleaved during reinstatement at consolidation (e.g., whilst asleep; Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010).

It should be noted that this episodic model of lexical storage is not inconsistent with speech perception models such as the distributed cohort model (DCM; Gaskell & Marslen-Wilson, 1997), and the perceived input/probe may still be conceived of as a multi-dimensional array[4]

15.3.2 Literature support for an episodic lexicon

The recent findings on pre-sleep lexical engagement are a problem for a complementary learning systems account of word learning because they imply that there are connexions between newly acquired words and words learnt long ago, with no mechanism for these abstract connexions to form so quickly. To the knowledge of the author, no formal model has yet been presented fully explaining the pre-sleep engagement data, despite three recent updates to CLSM (Kumaran et al., 2016; McClelland, 2013; McClelland et al., 2020). The updates to the CLSM allow for the integration of information rapidly when it overlaps with information already stored, following, for example, findings that schema-consistent information may be more efficiently learnt (Coutanche & Thompson-Schill, 2014; Havas et al., 2018; Tse et al., 2007). Although these updates may address the pre-sleep lexical *competition* findings (as paradigms measuring lexical competition require overlap between the novel competitor and the familiar target), they do not address other pre-sleep lexical *engagement* findings – for example, shifts in phoneme categorisation (Leach & Samuel, 2007; Lindsay et al., 2012), semantic associations (Bakker et al., 2015; Geukes et al., 2015; Tham et al., 2015) – or evidence of semantic retuning in response to recent experience (Rodd et al., 2016). Such findings are all inconsistent with a complementary learning systems account, since that account posits that words learnt long ago

[4]The central tenet of the DCM is that many representations are activated in parallel from fundamental information (e.g., particular frequencies) – forming a cohort of representations – which are then whittled down by the perception of further phonemes to a single lexical candidate. However, the model does not exclude the possibility that each lexical candidate may be represented by multiple stored episodes, each containing that candidate.

should be isolated in their own 'lexical' store, as abstract, generic representations (Davis & Gaskell, 2009; Lindsay & Gaskell, 2010).

However, work in the literature has demonstrated that words *may* have both episodic and abstract properties – strictly consistent only with an episodic account, such as that proposed by Goldinger (1998). Kapnoula and Samuel (2019) found that participants readily associated a particular referent and a novel word with a particular voice during training, and at test, were slower to identify the particular referent/word if it occurred with a different voice. Moreover, the researchers demonstrated that this effect was not modulated by sleep. Note that this last finding is in some ways similar to that shown by Experiments 5 and 6 (and Weighall et al., 2017) – the competition effect which was observed between novel and familiar words did not vary by whether the novel word was learnt on the same day, or the day before, testing.

Ostensibly, however, the episodic account is contradicted by another paper from Kapnoula and McMurray (2016a), again, reporting a pre-sleep effect. In that paper, Kapnoula and McMurray demonstrate that a novel word was still evoked, even if the voice at test was different from the voice heard during training. This implied exactly the opposite effect – abstraction away from the particular phonemic details of an episode. However, the episodic account does not posit that abstraction does not take place, merely that it takes place at a different point, and that what is *stored* is not abstract. Kapnoula and McMurray's finding of pre-sleep lexical competition *is* consistent with an episodic view of the lexicon, as stored traces from old and new words interacted immediately. Goldinger (1998) himself makes the point that evidence of 'speaker normalisation' (the ability to recognise words from different speakers, and generalise across them between training and testing) does not disprove his episodic account.

One might object that the episodic account is overly complicated or unparsimonious – particularly given the redundancy and multitude of traces. However, this seems unfair – the account is simple, as it argues only that what is heard is then put into memory, and then recalled (allowing for some degree of forgetting). It is however a weakness of both episodic and the complementary learning system accounts that they rely on abstraction processes which cannot be readily or directly observed in human participants (being derived from computational models which may or may not be biologically realistic). That said, an episodic account of the lexicon has an advantage over the abstractionist accounts (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010) insofar as it does not rely on abstraction occurring at a specific time point (e.g., during the hours of sleep Dumay & Gaskell, 2007), or under certain conditions (e.g., where new information is schema-consistent McClelland, 2013; McClelland et al., 2020). Instead, by an episodic account, abstraction is a consequence of the many echoes being represented across a set of output units.

15.3.3 Thesis findings in the context of an episodic lexicon

Goldinger's (1998) episodic account appears to explain the data from this thesis well. However, one must also consider how novel words are processed during lexical engagement, and for this, one must look to speech perception models, such as the

DCM (Gaskell & Marslen-Wilson, 1997). The DCM predicts that prior to the point of recognition, there is a 'lexical blend' of representations, activated according to their frequency in a listener's previous experience, and the degree to which they match the perceived input (e.g., a heard word). This 'blend' is technically termed a cohort. Consider a phonological competition trial in a mouse tracking task, with the referents CANDLE and CANDY on screen, and the input /kænd/ perceived (i.e., the portion of the word *before* the disambiguation point). At this point, the activation of both objects would be strong, and approximately equal, as they are both potential referents for /kænd/. Therefore, the ratio between these activations will be low (i.e., close to 1:1). This is detected as lexical competition, with participants being unable to decide between the two alternatives.

By contrast, on a perceptual competition trial with CANDLE and PARSNIP on screen, when the first phoneme /k/ is perceived, the ratio between the activation strength of both objects will be high – for the sake of the argument, say, 10:1, in favour of CANDLE. In this situation, the DCM predicts that the activation strength of one of the lexical candidates (here, CANDLE) is above threshold, and it's activation smothers the activation of the other candidate (here, PARSNIP, activated by its on-screen referent). Note that the DCM does not factor in inhibition – CANDLE does not inhibit PARSNIP, only overcomes it.

However, this account presents a problem when thinking about novel word lexical engagement. Consider a novel word phonological competition trial, in which ALIEN and ALIET are on screen and the participants hears /eɪl—/. The representation strength of the novel word must be low, because the activation strength is partially dependent upon frequency of occurrence and the novel competitor has been heard many fewer times than the familiar word. Why then does the familiar word's activation not smother that of the novel word? In short, in some studies, it does – this would explain why previous research has not shown evidence of lexical engagement until after sleep (where sleep purportedly strengthen and stabilises the novel word representations, giving rise to increases in lexical configuration performance, e.g., Dumay & Gaskell, 2007)

However, in studies reporting pre-sleep lexical engagement, something must either be boosting the novel word's activation, or decreasing the familiar word's activation, bringing the ratio of their strengths closer to 1:1. It seems likely that both of these effects are present. An episodic account is able to provide an explanation for why such effects occur, as discussed below.

Boosting the activation of the novel competitor

Consider learning the word 'aliet', as in Experiments 5–7 (and Weighall et al., 2017). The computation model which Goldinger (1998) uses in his episodic account of the lexicon, MINERVA 2 (Hintzman, 1986, 1988), would model this as follows. During training, participants were exposed to each novel word 12 times. Therefore, at least 12 novel word traces were stored – one for each exposure (possibly more, given that on some trials the participant repeated the novel word aloud after hearing it). Subsequently during mouse tracking, on the phonological competition trials, when a participant heard 'alien', with the referents ALIEN and ALIET on screen, the stored

novel word traces would each echo back with high intensity. This is because the voice used was the same across training and testing, and so the stored representations matched the heard word up until the disambiguation point. Furthermore, the same ALIET referent was present on screen during both training and testing. This all has the effect of boosting the activation of ALIET, as a potential referent when participants heard the word 'alien' (up until its disambiguation point). Although the novel object label had been heard only a small number times before during training, all of these exposures were in the exact same setting that participants found themselves in during testing. The high degree of overlap between the stimuli at training and testing (e.g., same speaker, same referent), as well as other quite banal similarities, such as participants being in the same room (cf., Pufahl & Samuel, 2014), may have compensated for the low number of exposures. In the vocabulary of Goldinger (1998), the intensity of the echo was high.

Decreasing the activation of the target

The factors that favoured the activation of the novel word would also have had the effect of reducing the activation of the familiar word. Although participants would have past episodic experience of both the word 'alien', and the concept ALIEN, the trial on which it was the target was the first time that participants had heard that particular speaker uttering 'alien', and it was the first time that participants had seen that particular ALIEN exemplar before. To identify 'alien', participants could not rely on any previous episode having a near complete overlap with the test trial – they instead needed to perform an abstraction from previous episodes. Presumably, this comes with a computational cost, and whilst this processing was taking place, the activation of 'alien' was decreased. The design of the experiment compounded this effect: each object only appeared on a single trial, forbidding the possibility of a recent episode of the target.

Furthermore, the target words in Experiments 5 and 6 have a low frequency of usage in everyday English[5]. Goldinger (1998) notes that a consequence of his model is that episodic effects (e.g., prior familiarity or otherwise with a speaker's voice) are particularly strong for low frequency words. This is because with a plethora of episodes from their experience to draw on, the probability of having an episode stored that more closely matches incoming perceptual input (e.g., of a heard word) is higher. For example, participants might have previously met a speaker with a similar sounding voice. With relatively few episodes to draw on, due to the target's low frequency of occurrence, the probably of a good match was lower.

It seems that under such experimental conditions, the activation of the target would have been particularly low, as the effect of being unfamiliar with the speaker's voice and the target exemplar would have been particularly strong – further enabling interference from the novel competitor, e.g., ALIET. The activation of ALIET, however, would have been higher than it otherwise could have been – given the relatively small number of exposures to the word. Combined, these two effects would bring the activation ratio closer to 1:1, and these two factors combined may have allowed the lexical competition effects to emerge.

[5]The average corpus frequency was reported by Weighall et al. (2017) to be ~ 8 per million

15.3.4 A hybrid model? Abstraction and consolidation in an episodic lexicon

Although a good fit for the data from Experiments 5 and 6, an episodic account does have some problems. The first problem is that although the capacity of long term memory is massive (e.g., Brady, Konkle, Alvarez & Olivia, 2008), processing all episodic traces of a word each time it is activated (in production or perception, as predicted by MINERVA 2; Goldinger, 1998; Hintzman, 1986, 1988) seems extremely inefficient and computationally unrealistic. Consider the high frequency of the word 'the' – this thesis alone contains over 4000 instances of it. Without any mechanism of abstraction across each context, all of these instances must be stored and processed separately. The computational costs of reading and writing these episodic traces to memory, and then processing them further during recognition, would be very high, perhaps unrealistically so.

The second problem is that there is some vagueness in what constitutes 'an episode', and how these may be bound together into more cohesive experiences (e.g., in order to remember what happened over longer periods of time). The way to solve this problem is to introduce a process like 'consolidation'. A consolidation-type process could amalgamate episodes into cohesive wholes, and for long term, more efficient storage, slowly abstract across the commonalities between episodes. Likewise, consolidation may function to stabilise representations, and embed them into cognitive networks (cf., Carpenter & Grossberg, 1988). This would explain the overnight improvements in stem completion performance from Experiments 5 and 6, as more stable and somehow strengthened traces, embedded into the rest of the network, would be easier for the cognitive system to retrieve.

To fit with an episodic account, however, this consolidation process would not be responsible for abstracting across recent experiences[6]. Stored traces must be represented as episodic and non-abstract in order to account for indexical effects (e.g., Kapnoula & Samuel, 2019). The proposal then would be to redefine consolidation for it to be more like how it was as it was originally conceived in the CLSM (McClelland et al., 1995), and remove some theory that word learning researchers have added to it (e.g., Davis & Gaskell, 2009; Lindsay & Gaskell, 2010). For example, to explain how episodic traces (stored in the hippocampus) do not engage on the first day of learning, Davis and Gaskell (2009) factored in a negative weighting for the hippocampal route, with the result being a system that favours cortical (i.e., abstract representation) processing. This addition should be removed, as episodic traces are readily accessible (e.g., Bowers et al., 2005; Kapnoula & Samuel, 2019; Pufahl & Samuel, 2014; Rodd et al., 2016). Additionally, in the original formulation of the CLSM, McClelland et al. (1995) discuss consolidation (and the abstraction resulting from it) potentially occurring over decades, and not as word learning researchers have suggested, overnight (e.g., Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Lindsay & Gaskell, 2010)[7]. Episodic representations would drive behaviour in the

[6]What constitutes 'recent', is an open question – there is a dearth of studies tracking representations over time (though e.g., Tamminen & Gaskell, 2008), and none which also show immediate lexical engagement effects.

[7]Note that this is not to suggest that these researchers believe that consolidation is completed

short term whilst consolidation acts to abstract experiences over longer time spans. Eventually, however, the computational demands of processing so many episodes would suggest that abstraction must take place, and there be a switch to behaviour driven by more abstract representations.

However, how might we explain the increase in stem completion performance overnight observed in Experiments 5 and 6? Assuming that this behaviour is driven by the same representations that drive lexical engagement effects – which are present immediately, and contain indexical information (Kapnoula & Samuel, 2019) – consolidation might act on the representations in their entirety – and not, as has been suggested, abstract away indexical details. By this account, even irrelevant details would initially be consolidated.

In summary, it is proposed that an episodic account needs a process like consolidation. Whilst in the short term, behaviour would be driven by episodic traces, in the long term, there needs to be a shift towards abstract representations. Consolidation would perform this function in the manner described by the original CLSM: the very slow and gradual interleaving of new and old information (McClelland et al., 1995). By contrast, in the short term, the function of consolidation would be to stabilise episodic representations, and to stitch them together to represent longer periods of time – but not abstract across them. At this stage, abstraction would occur as described by Goldinger (1998) – at word recognition processing, and not at storage. This would explain why indexical effect sizes do not appear to decrease with sleep (Kapnoula & Samuel, 2019), why in Experiments 5 and 6 (and in Weighall et al., 2017) one night of sleep did not make a difference to lexical engagement[8], and yet also why there are increases in lexical configuration performance.

It should be noted that, more recently, exactly the sort of mechanisms suggested above have been adopted by proponents of an episodic lexicon. The purpose of Goldinger's (1998) paper was to argue the case for an episodic lexicon, and for simplicity, the point was best made with a purely episodic computational model, MINERVA 2 (Hintzman, 1986, 1988). However, Goldinger (2007, p. 54) has revised his opinion, and writes:

> "The abstract lexicon is required to interpret [unusual input]; episodic memory is required to both generalise and delimit the effect. The [CLSM] offers a rapprochement for abstract and episodic theories of language; both forms of representation are mutually created in a reciprocal loop, uniting long-term memory with real-time perception."

There therefore seem to be agreement in the literature that an optimal theory would include a process like consolidation in an episodic account (see also Pierrehumbert, 2016).

overnight – they do not – cf., Tamminen and Gaskell (2008).

[8]Recall that words learnt yesterday and today were statistically indistinguishable on the most sensitive mouse tracking measure, path length. Weighall et al. (2017) showed the same effect in eye tracking.

15.4 Future work

The following section contains some brief thoughts about questions which cannot be answered with the current data, and are not otherwise addressed in the literature.

The most important continuation of this thesis is to run Experiment 7. Given that participants appear to have represented semantic knowledge, and that this was accessible whilst performing the lexical engagement task in Experiment 6, it seems likely that Experiment 7 would also find evidence that semantic representations support the lexical competition effects in Experiments 5 and 6. Assuming that it is the case that very recently learnt novel word representations contain semantic information, the following gives some ways that this thesis could be built upon. These ideas all concern how other aspects of a lexical representation may be abstract and generalisable, or episodic.

In many ways, learning a word is similar to learning a concept. When a child learns to apply the form 'cat' to the family pet, they must also learn that this form does not apply only to that specific individual animal. Instead, the form applies to a class containing all cats, which may vary across a number of properties (e.g., size, shape, colour). One area for future research would be to explore whether novel words that exhibit immediate lexical engagement also display evidence of semantic generalisation across different exemplars. This could be tested using the lexical engagement paradigm used in Experiments 5–7. For example, would changing some property of the referent between training and testing (e.g., its colour, size, shape or orientation) alter the lexical engagement effect?

Developmental data suggest that children were able to generalise immediately from a learnt exemplar to another from the same class. Holland, Mather, Simpson and Riggs (2016) showed that 3 and 4 year old children who learnt to associate an object label with a particular referent systematically extended this information to a different exemplar from the same class of objects. More impressively, this was under FM conditions, with only a single exposure. This finding suggests that in studies like Experiments 5 and 6, where participants explicitly learnt an object label with many more exposures, participants would also have no trouble in generalising across learnt and altered referents. This might manifest during lexical processing as a preserved competition effect, despite a change at test to some property of the novel competitor's referent.

On the other hand, an episodic account would suggest that changing the referent across training and testing sessions would reduce the size of the lexical competition effect. This is because, at test, the probe of episodic memory would contain the altered and on-screen referent, which would not be the same exemplar that participants had experience of during training. The intensity of the echo would therefore be reduced, with an associated weakening in the activation level of the novel competitor (cf., Goldinger, 1998). This reduction might then be sufficient to allow the novel competitor's activation to be smothered by the activation of the familiar target, thus depressing competition.

This extension would be important as it would not just allow one to test the claims made above with respect to an episodic lexicon, and episodic representations, but also give some insight into the relative time courses of processing of

stored semantic and phonological information. That they are different has been suggested in the literature (e.g., Coutanche & Thompson-Schill, 2014), but it is hard to understand why this would truly be the case, if novel word representations are unitary bindings between semantic and phonological information, as suggested by Experiment 6.

In a similar vein, Kapnoula and McMurray (2016a) reported that participants were able to abstract across different speakers between training and testing sessions, with an auditory mismatch paradigm, which allowed for the very specific activation of a competing novel form. However, would this finding replicate with stimuli and a design such as that used in Experiments 5 and 6? Recall that in these experiments, the activation of the novel competitor was less targeted. In Experiments 5 and 6, the novel word's activation was driven only by an on-screen novel referent and a stem shared with a familiar word – only the novel referent specifically evoked the novel form, as the stem was shared, and also activated the familiar word. By contrast, in the auditory mismatch paradigm, participants are played a word made from splicing a novel word form token and a familiar word form token together (e.g., 'jod' and 'job', being spliced together to form 'jo^{d}b' – with co-articulatory information on the vowel evoking the novel word). With a manipulation so specifically designed to evoke the novel word, the change of voices between training and testing may have had no effect, due to the very specific activation of the novel form permitted by the paradigm. With the novel representation so strongly activated, the ratio of the activation strengths would have remained low. However, the episodic account would suggest that anything that reduced the similarity between training and testing should cause an accompanying reduction in the size of the lexical competition effect observed. Therefore, if it was found to be the case that Kapnoula and McMurray's generalisation observation did not extend to mouse tracking, this would support an episodic account of the lexicon.

Finally, if in the above cases, generalisation was not observed (either across different speakers or referents at training/testing), one could seek to promote abstraction as a way of remedying this. Again, sticking with a design like that used in Experiments 5 and 6, this might be done by providing more diverse inputs during training. For example, if participants were trained with several voices and/or referents, would it be the case that this more generalised training resulted in a better ability to generalise at test? If so, this could still be explained in the context of an episodic account – with more and different episodes to draw on, there would be a higher probability that one approximates what is shown at test closely enough to boost the activation of the novel object sufficiently to permit it to engage in lexical competition. This fits with the views of Goldinger (2007), who argued that episodic memory permitted generalisation.

15.5 Conclusions

This thesis set out to investigate the most human of human capacities, language. Specifically, the aim was to investigate the nature of lexical representations, the building blocks of language.

The process of word learning may be described as the translation of remote in-

formation into associated and unified representations. This translation was thought to happen slowly. However, recent findings from the literature, and from those reported in this thesis, suggest it may take place rapidly. These data in turn raise questions such as how we should define a word, and when a lexical representation should be considered 'word-like'.

Such fundamental questions about how words are learnt, represented and stored cannot be answered in a single piece of work. Nevertheless, this thesis endorses an emerging view, found elsewhere in the literature, that old distinctions between 'lexical' and 'episodic' are no longer valid (e.g., Kapnoula & Samuel, 2019).

APPENDIX

A

APPENDIX FOR EXPERIMENTS 1 AND 2

Figure A.1: An example training trial from Experiment 1

Table A.1 Experiments 1 and 2 words

List 1 words		**List 2 words**		**Filler words**	
Familiar word	*Novel competitor*	*Familiar word*	*Novel competitor*		
amazon	alazon	anchor	amchor	badger	basket
bamboo	balboo	banana	banara	beetle	cabbage
celery	cedery	cradle	cragle	cavern	curtain
coffin	colfin	fossil	fostil	diesel	hamper
fabric	fablic	galaxy	ganaxy	kidney	lizard
parcel	pargel	garlic	garnic	loafer	monkey
pillar	piltar	guitar	guitur	needle	nettle
sleeve	sleere	helmet	holmet	orange	paddle
tarmac	talmac	jersey	jergey	peanut	pebble
tattoo	tartoo	meadow	mearow	pencil	raisin
violin	viodin	mosaic	motaic	tavern	tendon
walnut	walnot	potato	polato	tomato	turnip

APPENDIX

B

APPENDIX FOR EXPERIMENT 3

Figure B.1: SANDWICH, an example item used in Experiment 3

Table B.1 Experiment 3 words, and their sourcing

Experimental items					
Competing item one	**Competing item two**	**Distractor item**	**Labelling items**		
padlock[1]	paddle[1]	football[3]	headphones[3]	stapler[3]	table[3]
sandwich[1]	sandal[1]	helmet[3]	teabag[3]	needle[1]	bottle[3]
pencil[1]	penny[1]	hammer[3]	biscuit[1]	lampshade[3]	shower[3]
pasta[1]	pasty[1]	goggles[3]	mousetrap[3]	bucket[3]	lantern[1]
button[1]	butter[1]	speaker[3]	oven[3]	keyboard[3]	fountain[1]
towel[2]	tower[2]	scissors[3]	jacket[1]	banjo[3]	pillow[3]
parsnip[1]	parcel[3]	anchor[3]	ruler[3]	trombone[1]	sofa[3]
dolphin[2]	dollar[2]	guitar[1]	window[1]	battery[3]	bracelet[1]

Note. [1] denotes item is taken from Weighall et al. (2017), [2] denotes item is taken from Spivey et al. (2005), [3] denotes item was chosen by the experimenter.

APPENDIX

C

APPENDIX FOR EXPERIMENTS 4

Figure C.1: CAMPER, an example cartoon item used in Experiment 4

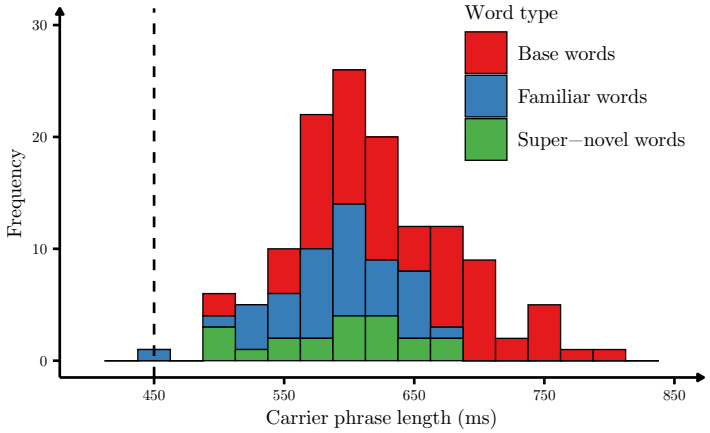


Figure C.2: Histogram of the length of the carrier phrases in Experiment 4–7's sound files. Vertical line at $x = 450$ms

Table C.1 Experiment 4 words

Competition	**Stimuli**			
	Cartoon		*Photo*	
	Target	*Competitor*	*Target*	*Competitor*
Perceptual	angel	mitten	battery	guitar
	baboon	flower	football	stapler
	beetle	diamond	goggles	biscuit
	candle	apple	hammer	bottle
	lantern	dragon	headphones	needle
	mermaid	french horn	helmet	trumpet
	monkey	rainbow	oven	scissors
	onion	chicken	pillow	anchor
	pumpkin	giraffe	speaker	bucket
	walrus	mirror	table	banjo
Phonological	bacon	baker	butter	button
	camel	camper	catalogue	caterpillar
	cartoon	carton	dollar	dolphin
	circle	circus	letter	lettuce
	kitchen	kitten	paddle	padlock
	lolly	lorry	parcel	parsnip
	medal	metal	pasty	pasta
	packet	package	penny	pencil
	robin	robber	sandal	sandwich
	window	winner	tower	towel

Note. Only the words in the columns labelled 'Target' were heard. Arrangement in the table does not denote spatial arrangement on-screen: left and right presentations were counterbalanced. *Cartoon* items from Weighall et al. (2017); *Photo* items from Experiment 3.

APPENDIX

D

APPENDIX FOR EXPERIMENTS 5, 6, AND 7

Table D.1 Experiment 5 (List 1) base and novel words

List 1					
Base word	*Novel word*	*Novel word IPA transcription*	*Base word*	*Novel word*	*Novel word IPA transcription*
alien	aliet	/eɪliət/	lantern	lantobe	/læntoʊb/
apricot	apricam	/eɪpɹɪkæm/	mayonnaise	mayonnote	/meɪənoʊt/
baboon	baboop	/bæbuːp/	napkin	napkig	/næpkɪg/
bikini	bikinar	/bɪkinɑː/	ornament	ornameld	/ɔːnəmɛld/
bracelet	bracelop	/bɹeɪsləp/	parade	parafe	/pəɹeɪf/
cactus	cactul	/kæktəl/	potato	potatuck	/pəteɪtʊk/
caramel	caramen	/kæɹəmɛn/	pumpkin	pumpkige	/pʊmpkɪʒ/
chimpanzee	chimpantu	/tʃɪmpæntuː/	rugby	rugbock	/ɹʊgbɐk/
dolphin	dolphik	/dɐlfɪk/	skeleton	skeledu	/skɛlədu:/
donkey	donkop	/dɐŋkɐp/	squirrel	squirrome	/skwɪɹoʊm/
fountain	fountel	/faʊntəl/	tissue	tissove	/tɪʃoʊv/
graffiti	graffino	/gɹəfiːnəʊ/	walnut	walnog	/wɔːlnɐg/

Table D.2 Experiment 5 (List 2) base and novel words

List 2					
Base word	*Novel word*	*Novel word IPA transcription*	*Base word*	*Novel word*	*Novel word IPA transcription*
angel	angesh	/eɪndʒiːʃ/	nugget	nuggev	/nʊgəv/
badminton	badminteef	/bædmɪn-tiːf/	onion	oniot	/ʊniːət/
biscuit	biscal	/bɪskəl/	pelican	pelical	/pɛlikəl/
bramble	brambo	/bɹæmboʊ/	penguin	pengwove	/pɛŋgwoʊv/
broccoli	broccaroo	/bɹɐkəɹuː/	pyramid	pyramin	/pɪɹəmɪn/
caravan	caravat	/kæɹəvæt/	sergeant	sergeast	/sɑːdʒiːst/
chocolate	chocolor	/tʃɐkəlɔː/	signature	signatik	/sɪgnətɪk/
costume	costuke	/kɐstjuːk/	somersault	somersaumf	/sʊməsɔːmf/
daffodil	daffodote	/dæfədoʊt/	target	targil	/tardʒɪl/
dinosaur	dinosut	/dainəsʊt/	tattoo	tattefe	/tætiːf/
gadget	gadgel	/gædʒəl/	trombone	trombal	/trɐmbəl/
mermaid	mermiff	/məːmɪf/	walrus	walrick	/wɔːlɪɪk/

Table D.3 Experiment 5 (List 3) base and novel words

List 3					
Base word	*Novel word*	*Novel word IPA transcription*	*Base word*	*Novel word*	*Novel word IPA transcription*
athlete	athlove	/æθloʊv/	mushroom	mushrood	/mʊʃɹuːd/
balcony	balcozo	/bælkəzoʊ/	octopus	octopum	/ɐktəpʊm/
blossom	blossail	/blɐseɪl/	parachute	parasheff	/pæɹəʃɛf/
breakfast	breakfal	/bɹɛkfəl/	parsnip	parsnin	/pɑːsnɪn/
buffalo	buffaluk	/bʊffəlʊk/	picnic	picnin	/pɪknɪn/
cardigan	cardigite	/kɑːdɪgaɪt/	reptile	reptite	/ɹɛptaɪt/
clarinet	clarinone	/klæɪnoʊn/	siren	siredge	/saɪɹɪdʒ/
crocodile	crocodol	/kɹɐkədol/	spider	spidet	/spaɪdɛt/
dungeon	dungeoth	/dʊndʒəθ/	tornado	tornadus	/tɔːneɪdəs/
flamingo	flamingist	/flæmi-ŋgɪst/	tulip	tulode	/tjuːloʊd/
guitar	guitas	/gɪtæs/	volcano	volcagi	/eɪliət/
kangaroo	kangami	/kæŋgəmiː/	yoghurt	yogem	/jɐgəm/

Table D.4 Experiment 5 sound file properties in milliseconds for those items used in the lexical engagement task

Property	**Statistic**	**Stimuli type**		
		Base word	*Familiar*	*Super-novel*
Total length	*M*	1683	1399	1558
	SD	142	115	119
Carrier phrase	*M*	638	588	590
	SD	65	45	50
Target label	*M*	1045	810	968
	SD	139	110	119
DP	*M*	1103	898	—
	SD	129	82	—

Note. DP = 'disambiguation point'; the point (measured from the start of the sound file) at which a target may be disambiguated from its competitor which may, or may not, have been present on any given trial. Consequently, super-novel words have no such point, as they did not have competitors.

Figure D.1: ANGESH, an example novel referent used in Experiments 5–7

Table D.5 Experiment 5 familiar and super-novel words

Familiar words		**Super-novel words**	
List A target	*List B target*	*Word*	*IPA transcription*
bacon	baker	balras	/bælɹæs/
beaker	beetle	chamgalp	/tʃæmgælp/
butter	button	grompa	/gɹɐmpə/
camper	camel	hekobi	/hɛkoʊbɪː/
candle	candy	hinshink	/hɪnʃɪŋk/
cartoon	carton	kipthermit	/kɪpθɛːmɪt/
caterpillar	catalogue	molsmit	/mɐlsmɪt/
circle	circus	nemok	/niːmɐk/
kitten	kitchen	nishboka	/nɪʃboʊkə/
letter	lettuce	nolcrid	/nɐlkɹɪd/
lolly	lorry	shoboe	/ʃoʊboʊ/
medal	metal	sloskonad	/slɐskɐnæd/
monkey	money	snidfey	/snɪdfiː/
packet	package	stansert	/stænzɛːt/
paddle	padlock	sunipog	/sʊniːpɐg/
pasta	pasty	tastanza	/tæstænzə/
pencil	penny	tegwop	/tɛgwɐp/
robin	robber	trolkey	/tɹɐlkiː/
sandal	sandwich	twamket	/twæmkɛt/
window	winner	vopum	/voʊpʊm/

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Anderson, S. E. & Spivey, M. J. (2009). The enactment of language: decades of interactions between linguistic and motor processes. *Language and Cognition*, *1*(1), 87–111. doi: 10.1515/langcog.2009.005

Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, *28*(02), 105–167. doi: 10.1017/S0140525X05000038

Atir-Sharon, T., Gilboa, A., Hazan, H., Koilis, E. & Manevitz, L. M. (2015). Decoding the formation of new semantics: MVPA investigation of rapid neocortical plasticity during associative encoding through fast mapping. *Neural Plasticity*, *2015*, 804385. doi: 10.1155/2015/804385

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behaviour Research Methods*, *37*(3), 379–384. doi: 10.3758/bf03192707

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G. & McQueen, J. M. (2015). Tracking lexical consolidation with ERPs: lexical and semantic-priming effects on N400 and LPC responses to newly-learned words. *Neuropsychologia*, *79*, 33–41. doi: j.neuropsychologia.2015.10.020

Barca, L. & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLoS ONE*, *7*(4), e35932. doi: 10.1371/journal.pone.0035932

Barca, L. & Pezzulo, G. (2015). Tracking second thoughts: continuous and discrete revision processes during visual lexical decision. *PLoS ONE*, *10*, e0116193. doi: 10.1371/journal.pone.0116193

Barr, D. J. & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441–455. doi: 10.1080/01690960903047122

Bartolotti, J. & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129–1147. doi: 10.1111/j.1551-6709.2012.01243.x

Bartolotti, J., Schroeder, S. R., Hayakawa, S., Rochanavibhata, S., Chen, P. & Marian, V. (2020). Listening to speech and non-speech sounds activates phonological and semantic knowledge differently. *Quarterly Journal of Experimental Psychology*, *73*(8), 1135–1149. doi: 10.1177/1747021820923944

Betts, H. N., Gilbert, R. A., Cai, Z. G., Okedara, Z. B. & Rodd, J. M. (2017). Retuning of lexical-semantic representations: repetition and spacing effects in word-meaning representation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *44*(7), 1130–1150. doi: 10.1037/xlm0000507

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R. & Bendayan, R. (2017). Non-normal data: is ANOVA still an option? *Psicothema*, *29*(4), 552–557. doi: 10.7334/psicothema2016.383

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9), 341–345.

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*, B45–B54. doi: 10.1016/j.cognition.2005.02.002

Brady, T. F., Konkle, T., Alvarez, G. A. & Olivia, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. doi: 10.1073/pnas.0803390105

Brown, H., Weighall, A., Henderson, L. M. & Gaskell, M. G. (2012). Enhanced recognition and recall of new words in 7- and 12-year-olds following a period of offline consolidation. *Journal of Experimental Child Psychology*, *112*(1), 56–72. doi: 10.1016/j.jecp.2011.11.010

Cabeza, R., Kapur, S., Craik, F. I. M., Houle, S. & Tulving, E. (1997). Functional neuroanatomy of recall and recognition: a PET study of episodic memory. *Journal of Congitive Neuroscience*, *9*(2), 254–265. doi: 10.1162/jocn.1997.9.2.254

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

Calcagnì, A., Lombardi, L. & Sulpizio, S. (2017, December). Analyzing spatial data from mouse tracker methodology: an entropic approach. *Behavior Research Methods*, *49*(6), 2012–2030. doi: 10.3758/s13428-016-0839-5

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, *6*(3), 184–205. doi: 10.1080/15475441.2010.484379

Carey, S. & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17-29.

Carpenter, G. A. & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer: Special Issue on Artificial Neural Systems*, *21*(3), 77–88. doi: 10.1109/2.33

Carreiras, M., Perea, M. & Grainger, J. (1997). Effects of the orthographic neighborhood in visual word recognition: cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*(4), 857–871. doi: 10.1037/0278-7393.23.4.857

Chomsky, N. (1959). Review of Verbal Behavior. *Language*, *35*(1), 26-58. doi: 10.2307/411334

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). United States of America: Lawrence Erlbaum Associates. doi: 10.4324/

9780203771587

Cooper, E., Greve, A. & Henson, R. N. (2019a). *Fast mapping (FM) in adults: a proposed attempt to replicate evidence from implicit memory.* Pre-registration poster. Retrieved from https://osf.io/9wqfg/

Cooper, E., Greve, A. & Henson, R. N. (2019b). Investigating fast mapping task components: no evidence for the role of semantic referent nor semantic inference in healthy adults. *Frontiers in Psychology*, *10*(394), 1–12. doi: 10.3389/fpsyg.2019.00394

Cooper, E., Greve, A. & Henson, R. N. (2019c). Little evidence for fast mapping (FM) in adults: a review and discussion. *Cognitive Neuroscience*, *10*(4), 196–209. doi: 10.1080/17588928.2018.1542376

Cooper, E., Greve, A. & Henson, R. N. (2019d). Response to commentaries on our review of fast mapping in adults. *Cognitive Neuroscience*, *10*(4), 237–240. doi: 10.1080/17588928.2019.1651709

Coutanche, M. N. (2019). Addressing misconceptions of fast mapping in adults. *Cognitive Neuroscience*, *10*(4), 226–228. doi: 10.1080/17588928.2019.1593955

Coutanche, M. N. & Koch, G. E. (2017). Variation across individuals and items determine learning outcomes from fast mapping. *Neuropsychologia*, *106*, 187–193. doi: 10.1016/j.neuropsychologia.2017.09.029

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

Coutanche, M. N. & Thompson-Schill, S. L. (2015). Rapid consolidation of new knowledge in adulthood via fast mapping. *Trends in Cognitive Sciences*, *19*(9), 486–488. doi: 10.1016/j.tics.2015.06.001

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. doi: 10.1037/0096-3445.104.3.268

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. doi: 10.1080/01690960143000074

Dale, R. & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*(5), 983–996. doi: 10.1111/j.1551-6709.2010.01164.x

Dale, R., Kehoe, C. & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition*, *35*(1), 15–28. doi: 10.3758/BF03195938

Dale, R., Roche, J., Snyder, K. & McCall, R. (2008). Exploring action dynamics as an index of paired-associate learning. *PLoS ONE*, *3*(3), e1728. doi: 10.1371/journal.pone.0001728

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*(1536), 3773–3800. doi: 10.1098/rstb.2009.0111

Dilkina, K., McClelland, J. L. & Plaut, D. C. (2010). Are there mental lexicons?

the role of semantics in lexical decision. *Brain Research*, *1365*, 66–81. doi: 10.1016/j.brainres.2010.09.057

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. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition*, *123*(1), 119–132. doi: 10.1016/j.cognition .2011.12.009

Dumay, N., Gaskell, M. G. & Feng, X. (2004). A day in the life of a spoken word. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26), 339–344.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B. & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychology Methods*, *1*(2), 170–177. doi: 10.1037/1082-989x.1.2.170

Duran, N. D., Dale, R. & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin and Review*, *17*(4), 486–491. doi: 10.3758/pbr.17.4.486

Dysart, E. L., Mather, E. & Riggs, K. J. (2016). Young children's referent selection is guided by novelty for both words and actions. *Journal of Experimental Child Psychology*, *146*, 231–237. doi: 10.1016/j.jecp.2016.01.003

Ebbinghaus, H. (1913). *Memory: a contribution to experimental psychology* (C. E. Bussenius, Trans.). Teachers College Press. doi: 10.1037/10011-000

Elward, R. L., Dzieciol, A. M. & Vargha-Khadem, F. (2019). Little evidence for fast mapping in adults with developmental amnesia. *Cognitive Neuroscience*, *10*(4), 215–217. doi: 10.1080/17588928.2019.1593123

Farmer, T. A., Anderson, S. E. & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory and Language*, *57*(4), 570–595. doi: 10.1016/j.jml.2007.04.003

Farmer, T. A., Cargill, S., Hindy, N., Dale, R. & Spivey, M. (2007). Tracking the continuity of language comprehension: computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science*, *31*(5), 889–909. doi: 10.1080/03640210701530797

Feather, J., Vélez, N. & Saxe, R. (2014). *Replication of: "Action dynamics reveal parallel competition in decision making", by McKinstry, Dale, and Spivey (2008, Psychological Science).* Retrieved from https://osf.io/d0n81/

Fechner, G. T. (1860/1966). *Elements of psychophysics* (Vol. 1; D. H. Howes & E. G. Boring, Eds.). New York City, New York: Holt, Rinehart and Winston.

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

Forster, K. I. & Forster, J. C. (2003). DMDX: a Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, *35*(1), 116–124. doi: 10.3758/BF03195503

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315–323. doi: 10.1177/0963721417746793

Freeman, J. B. & Ambady, N. (2009). Motions of the hand expose the partial and

parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183–1188. doi: 10.1111/j.1467-9280.2009.02422.x

Freeman, J. B. & Ambady, N. (2010). MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method. *Behaviour Research Methods*, *42*, 226–241. doi: 10.3758/brm.42.1.226

Freeman, J. B. & Ambady, N. (2011). Hand movements reveal the time-course of shape and pigmentation processing in face categorization. *Psychonomic Bulletin and Review*, *18*, 705–712. doi: 10.3758/s13423-011-0097-6

Freeman, J. B. & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, *45*(1), 83–97. doi: 10.3758/s13428-012-0225-x

Freeman, J. B., Dale, R. & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 1–6. doi: 10.3389/fpsyg.2011.00059

Gabrieli, J. D. E., Cohen, N. J. & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain and Cognition*, *7*(2), 157–177. doi: 10.1016/0278-2626(88)90027-9

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

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

Gaskell, M. G. & Lindsay, S. (2019). Reasons to doubt the generalizability, reliability, and diagnosticity of fast mapping (FM) for rapid lexical integration. *Cognitive Neuroscience*, *10*(4), 234–236. doi: 10.1080/17588928.2019.1600487

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. doi: 10.1080/016909697386646

Gernsbacher, M. A. & Morson, E. (2019). Fast mapping is a laboratory task, not a cognitive capacity. *Cognitive Neuroscience*, *10*(4), 223–225. doi: 10.1080/17588928.2019.1573810

Geukes, S., Gaskell, M. G. & Zwisterlood, P. (2015). Stroop effects from newly learnted color words: effects of memory consolidation and episodic context. *Frontiers in Psychology*, *6*(278), 1–17. doi: 10.3389/fpsyg.2015.00278

Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. doi: 10.5812/ijem.3505

Gilboa, A. (2019). Long-term fragility: interference susceptibility may be an inherent characteristic of memory traces acquired through fast mapping. *Cognitive Neuroscience*, *10*(4), 218–220. doi: 10.1080/17588928.2019.1593122

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. doi: 10.1037/0033-295X.105.2.251

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th International Congress of*

Phonetic Sciences (pp. 49–54). Saarbrücken, Germany.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M. & Wenger, N. R. (1992). Children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99–108. doi: 10.1037/0012-1649.28.1.99

Gow, D. W., Jr. & Olson, B. B. (2015). Sentential influences on acoustic-phonetic processing: a Granger causality analysis of multimodal imaging data. *Language, Cognition and Neuroscience*, *31*(7), 841–855. doi: 10.1080/23273798.2015.1029498

Greve, A., Cooper, E. & Henson, R. N. (2014). No evidence that 'fast-mapping' benefits novel learning in healthy older adults. *Neuropsychologia*, *60*, 52–59. doi: 10.1016/j.neuropsychologia.2014.05.011

Halberda, J. (2006). Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*(4), 310–344. doi: 10.1016/j.cogpsych.2006.04.003

Hartigan, J. A. & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70–84. doi: 10.1214/aos/1176346577

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

Hawkins, E., Astle, D. E. & Rastle, K. (2014). Semantic advantage for learning new phonological form representations. *Journal of Cognitive Neuroscience*, *27*(4), 775–786. doi: 10.1162/jocn_a_00730

Hawkins, E. & Rastle, K. (2016). How does the provision of semantic information influence the lexicalization of new spoken words? *Quarterly Journal of Experimental Psychology*, *69*(7), 1322–1339. doi: 10.1080/17470218.2015.1079226

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B. & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychology and Personality Science*, *5*(7), 815–824. doi: 10.1177/1948550614534701

Hehman, E., Stolier, R. M. & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes and Intergroup Relations*, *18*(3), 384–401. doi: 10.1177/1368430214538325

Henderson, L. M. & James, E. (2018). Consolidating new words from repetitive versus multiple stories: prior knowledge matters. *Journal of Experimental Child Psychology*, *166*, 465–484. doi: 10.1016/j.jecp.2017.09.017

Henderson, L. M., Powell, A., Gaskell, M. G. & Norbury, C. (2014). Learning and consolidation of new spoken words in autism spectrum disorder. *Developmental Science*, *17*(6), 858–871. doi: 10.1111/desc.12169

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

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402. doi: 10.1038/nrn2113

Himmer, L., Müller, E., Gais, S. & Schönauer, M. (2017). Sleep-mediated memory

consolidation depends on the level of integration at encoding. *Neurobiology of Learning and Memory*, *137*(2017), 101–106. doi: 10.1016/j.nlm.2016.11.019

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). Judgements 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

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97. doi: /10.1038/scientificamerican0960-88

Holland, A., Mather, E., Simpson, A. & Riggs, K. (2016). Get your facts right: preschoolars systematically extend both object names and category-relevant facts. *Frontiers in Psychology*, *7*, 1–9. doi: 10.3389/fpsyg.2016.01064

Horst, J. S., Scott, E. J. & Pollard, J. A. (2010). The role of competition in word learning via referent selection. *Developmental Science*, *13*(5), 706–713. doi: 10.1111/j.1467-7687.2009.00926.x

Huettig, F., Rommers, J. & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologia*, *137*(2), 151–171. doi: 10.1016/j.actpsy.2010.11.003

Johnson, A., Mulder, B., Sijbinga, A. & Hulsebos, L. (2012). Action as a window to perception: measuring attention with mouse movements. *Applied Cognitive Psychology*, *26*(5), 802–809. doi: 10.1002/acp.2862

Kaminski, J., Call, J. & Fischer, J. (2004). Word learning in a domestic dog: evidence for "fast mapping". *Science*, *304*(5677), 1682–1683. doi: 10.1126/science.1097859

Kapnoula, E. C. & McMurray, B. (2016a). Newly learned word forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin and Review*, *23*(2), 491–499. doi: 10.3758/s13423-015-0897-1

Kapnoula, E. C. & McMurray, B. (2016b). Training alters the resolution of lexical interference: evidence for plasticity of competition and inhibition. *Journal of Experimental Psychology: General*, *145*(1), 8–30. doi: 10.1037/xge0000123

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

Kapnoula, E. C. & Samuel, A. G. (2019). Voices in the mental lexicon: words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111–127. doi: 10.1016/j.jml.2019.05.001

Kieslich, P. J. & Henninger, F. (2017). Mousetrap: an integrated, open-source mouse-tracking package. *Behaviour Research Methods*, *49*, 1652–1667. doi: 10.3758/s13428-017-0900-z

Kieslich, P. J., Schoemann, M., Grage, T., Hepp, J. & Scherbaum, S. (2020). Design factors in mouse-tracking: what makes a difference? *Beheavior Research Methods*, *52*(1), 317–341. doi: 10.3758/s13428-019-01228-y

Kieslich, P. J., Wulff, D. U., Henninger, F., Haslbeck, J. M. B. & Brockhaus, S. (2020). Package documentation for mousetrap (3.1.5 ed.) [Computer software manual].

Kim, J. J. & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, *256*(5057), 675–677. doi: 10.1126/science.1585183

Korenic, S. A., Nisonger, S. J., Krause, B. W., Wijtenburg, S. A., Hong, L. E. & Rowland, L. M. (2016). Effectiveness of fast mapping to promote learning in schizophrenia. *Schizophrenia Research: Cognition*, *4*, 24–31. doi: 10.1016/j.scog.2016.04.003

Koutstaal, W. (2019). Other 'routes in'? Has the 'fast' in the fast mapping concept led us astray? *Cognitive Neuroscience*, *10*(4), 213–214. doi: 10.1080/17588928.2019.1593124

Kumaran, D., Hassabis, D. & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7). doi: 10.1016/j.tics.2016.05.004

Laine, M., Polonyi, T. & Abari, K. (2013). More than words: fast acquisition and generalization of orthographic regularities during novel word learning in adults. *Journal of Psycholinguistic Research*, *43*(4), 381–396. doi: 10.1007/s10936-013-9259-1

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

Lindsay, S. & Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning*, *60*(2), 45–63. doi: 10.1111/j.1467-9922.2010.00600.x

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

Lindsay, S., Sedin, L. M. & 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. doi: 10.1016/j.jml.2011.07.005

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. doi: 10.1097/00003446-199802000-00001

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*, 151–169. doi: 10.1146/annurev.publheath.23.100901.140546

Magnuson, J. S. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences*, *102*(29), 9995–9996. doi: 10.1073/pnas.0504413102

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. doi: 10.1037/0096-3445.132.2.202

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

Maldonado, M., Dunbar, E. & Chemla, E. (2019). Mouse tracking as a window into decision making. *Behavior Research Methods*, *51*(3), 1085–1101. doi:

10.3758/s13428-018-01194-x

Marghetis, T., Núñez, R. & Bergen, B. K. (2014). Doing arithmetic by hand: hand movements during exact arithmetic reveal systematic, dynamic spatial processing. *Quarterly Journal of Experimental Psychology*, *67*(8), 1579–1596. doi: 10.1080/17470218.2014.897359

Markson, L. & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*, 813–815. doi: 10.1038/385813a0

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. doi: 10.1016/s0749-596x(02)00037-2

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210. doi: 10.1037/a0033812

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

McClelland, J. L., McNaughton, B. L. & Lampinen, A. K. (2020). Integration of new information in memory: new insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1799). doi: 10.1098/rstb.2019.0637

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

McKay, A., Davis, C., Savage, G. & Castles, A. (2008). Semantic involvement in reading aloud: evidence from a non-word training study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*(6), 1495–1517. doi: 10.1037/a0013357

McMurray, B., Kapnoula, E. C. & Gaskell, M. G. (2017). Learning and integration of new word-forms: Consolidation, pruning and the emergence of automaticity. In M. G. Gaskell & J. Mirković (Eds.), *Speech perception and spoken word recognition* (1st ed., pp. 116–143). 2 Park Square, Milton Park, Abingdon, Oxfordshire, OX14 4RN: Routledge. doi: 10.4324/9781315772110

McNeil, A. M. & Johnston, R. S. (2004). Word length, phonemic and visual similarity effects in poor and normal readers. *Memory and Cognition*, *32*(5), 687–695. doi: 10.3758/BF03195859

Merhav, M., Karni, A. & Gilboa, A. (2014). Neocortical catastrophic interference in healthy and amnesic adults: A paradoxical matter of time. *Hippocampus*, *24*(12), 1653–1662. doi: 10.1002/hipo.22353

Merhav, M., Karni, A. & Gilboa, A. (2015). Not all declarative memories are created equal: fast mapping as a direct route to cortical declarative representations. *NeuroImage*, *117*, 80–92. doi: 10.1016/j.neuroimage.2015.05.027

Mervis, C. B. & Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Development*, *65*(6), 1646–1662. doi: 10.1111/j.1467-8624.1994.tb00840.x

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of

words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi: 10.1037/h0031564

Nation, P. & Waring, R. (1997). Vocabularly size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6–19). The Edinburgh Building, Cambridge, CB2 2RU, United Kingdom: Cambridge University Press.

Navalpakkam, V. & Churchill, E. (2012). Mouse tracking: measuring and predicting users' experience of web-based content. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2963–2972. doi: 10.1145/2207676.2208705

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

O'Connor, R. J. & Riggs, K. J. (2019). Adult fast mapping memory research is based on a misconception of developmental word learning data. *Current Directions in Psychological Science*, *28*(6), 528–533. doi: 10.1177/0963721419858426

Oh, Y. M., Todd, S., Beckner, C., Hay, J., King, J. & Needle, J. (2020). Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific Reports*, *10*(22318), 1–9. doi: 10.1038/s41598-020-78810-4

O'Connor, R. J., Lindsay, S., Mather, E. & Riggs, K. J. (2019). Why would a special FM process exist in adults, when it does not appear to exist in children? *Cognitive Neuroscience*, *10*(4), 221–222. doi: 10.1080/17588928.2019.1574260

Palma, P. & Titone, D. (2020). Something old, something new: a review of the literature on sleep-related lexicalization of novel words in adults. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-020-01809-5

Palombo, D. J., Williams, L. J., Abdi, H. & Levine, B. (2013). The survey of autobiographical memory (SAM): a novel measure of trait memories in everyday life. *Cortex*, *49*(6), 1526–1540. doi: 10.1016/j.cortex.2012.08.023

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R. & Freeman, J. B. (2013). Good things peak in pairs: a note on the bimodality coefficient. *Frontiers in Psychology*, *4*, 1–4. doi: 10.3389/fpsyg.2013.00700

Pierrehumbert, J. B. (2016). Phonological representation: beyond abstract versus episodic. *Annual Review of Linguistics*, *2*, 33–52. doi: 10.1146/annurev-linguistics-030514-125050

Pinker, S. (1995). *The language instinct: the new science of language and the mind*. Penguin Books.

Pufahl, A. & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integration auditory memory representations. *Cognitive Psychology*, *70*, 1–30. doi: 10.1016/j.cogpsych.2014.01.001

Qiao, X., Forster, K. & Witzel, N. (2009). Is banara really a word? *Cognition*, *113*(2), 254–257. doi: 10.1016/j.cognition.2009.08.006

R Core Team. (2021). R: A language and environment for statistical computing (4.0.5 ed.) [Computer software manual]. Vienna, Austria. Retrieved from https://cran.r-project.org/doc/manuals/fullrefman.pdf

Richman, J. S. & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology*

– *Heart and Circulatory Physiology*, *278*(6), H2039–H2049. doi: 10.1152/ajpheart.2000.278.6.H2039

Riggs, K. J., Mather, E., Hyde, G. & Simpson, A. (2015). Parallels between action-object and word-object mappings in young children. *Cognitive Science*, *40*(2016), 992-1006. doi: 10.1111/cogs.12262

Rivas, E. (2005). Recent use of signs by chimpanzees (Pan troglodytes) in interactions with humans. *Journal of Comparative Psychology*, *119*(4), 404–417. doi: 10.1037/0735-7036.119.4.404

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

Rogers, T. T. & McClelland, J. L. (2014). Parallel distributed processing at 25: further explorations in the microstructure of cognition. *Cognitive Science*, *38*(6), 1024–1077. doi: 10.1111/cogs.12148

Rosenthal, R. (1994). Parametric measures of effect size. In *The Handbook of Research Synthesis* (pp. 231–244). Russell Sage Foundation.

Sakhon, S., Edwards, K., Luongo, A., Murphy, M. & Edgin, J. (2018). Small sets of novel words are fully retained after 1-week in typically developing children and down syndrome: a fast mapping study. *Journal of the International Neuropsychological Society*, *24*(9), 955–965. doi: 10.1017/S1355617718000450

Sandfeld, J. & Jensen, B. R. (2005). Effect of computer mouse gain and visual demand on mouse clicking performance and muscle activation in a young and elderly group of experienced computer users. *Applied Ergonomics*, *36*(5), 547–555. doi: 10.1016/j.apergo.2005.03.003

SAS Institute Inc. (2018). SAS/STAT 15.1 User's Guide [Computer software manual]. Cary, NC.. Retrieved from https://documentation.sas.com/

Schneider, W., Eschman, A. & Zuccolotto, A. (2002). Software manual for E-Prime 2.0 [Computer software manual]. Pittsburgh, PA.

Scoville, W. B. & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry*, *20*(1), 11–21. doi: 10.1136/jnnp.20.1.11

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*, *108*(3), 1146–1151. doi: 10.1073/pnas.1005238108

Skinner, B. F. (1957). *Verbal behavior*. Copley Publishing Group. doi: 10.1037/11256-000

Smith, C. N., Urgolites, Z. J., Hopkins, R. O. & Squire, L. R. (2014). Comparison of explicit and incidental learning strategies in memory-impaired patients. *Proceedings of the National Academy of Sciences*, *111*(1), 475-479. doi: 10.1073/pnas.1322263111

Smith, M. W., Sharit, J. & Czaja, S. J. (1999). Aging, motor control, and the performance of computer mouse tasks. *Human factors*, *41*(3), 389–396. doi: 10.1518/001872099779611102

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. doi: 10.1037/a0014509

Snowling, J., John, W., Adams, D. V. M., Bishop, S. & Stothard, M. (2001). Education attainments of school leavers with a preschool history of speech-language impairments. *International Journal of Language and Communication Disorders*, *36*(2), 173–183. doi: 10.1080/13682820120976

Song, J.-H. & Nakayama, K. (2008). Target selection in visual search as revealed by movement trajectories. *Vision Research*, *48*(7), 853–861. doi: 10.1016/j.visres.2007.12.015

Spivey, M. J. (2016). Semantics influences speech perception: commentary on Gow and Olson (2015). *Language, Cognition and Neuroscience*, *31*(7), 856–859. doi: 10.1080/23273798.2016.1140788

Spivey, M. J. & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. doi: 10.1111/j.1467-8721.2006.00437.x

Spivey, M. J., Grosjean, M. & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, *102*(29), 10393–10398. doi: 10.1073/pnas.0503903102

Squire, L. R. & Cohen, N. (1979). Memory and amnesia: resistance to disruption develops for years after learning. *Behavioral and Neural Biology*, *25*(1), 115–125. doi: 10.1016/S0163-1047(79)90841-0

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi: 10.1037/h0054651

Swingley, D. (2010). Fast mapping and slow mapping in children's word learning. *Language Learning and Development*, *6*, 179–183. doi: 10.1080/15475441.2010.484412

Szmalec, A., Page, M. P. A. & Duyck, W. (2012). The development of long-term lexical representations through Hebb repetition learning. *Journal of Memory and Language*, *67*(3), 342–354. doi: 10.1016/j.jml.2012.07.001

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

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. doi: 10.1523/jneurosci.3028-10.2010

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

Teyler, T. J. & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*(2), 147–154. doi: 10.1037/0735-7044.100.2.147

Tham, E. K. H., Lindsay, S. & Gaskell, M. G. (2015). Markers of automaticity in sleep-associated consolidation of novel words. *Neuropsychologia*, *71*, 146–157. doi: 10.1016/j.neuropsychologia.2015.03.025

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., ... Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*,

316(5821), 76–82. doi: 10.1126/science.1135935

Tulving, E. (1984). Précis of 'Elements of episodic memory'. *Behavioral and Brain Sciences*, *7*(2), 223–228. doi: 10.1017/S0140525X0004440X

van der Wel, R. P. R. D., Sebanz, N. & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133. doi: 10.1016/j.cognition.2013.10.004

Vlach, H. A. & Sandhofer, C. M. (2012). Fast mapping across time: memory processes support children's retention of learned words. *Frontiers in Psychology*, *3*(46), 1–8. doi: 10.3389/fpsyg.2012.00046

Wagner, R. K., Torgesen, J. K., Rashotte, C. A. & Pearson, N. A. (1999). *Comprehensive test of phonological processing.*

Walker, S., Henderson, L. M., Fletcher, F. E., Knowland, V. C. P., Cairney, S. A. & Gaskell, M. G. (2019). Learning to live with interfering neighbours: the influence of time of learning and level of encoding on word learning. *Royal Society Open Science*, *6*(4), 181842. doi: 10.1098/rsos.181842

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

Warren, D. E. & Duff, M. C. (2014). Not so fast: hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus*, *24*(8), 920–933. doi: 10.1002/hipo.22279

Warren, D. E. & Duff, M. C. (2019). Fast mappers, slow learners: word learning without hippocampus is slow and sparse irrespective of methodology. *Cognitive Neuroscience*, *10*(4), 210–212. doi: 10.1080/17588928.2019.1593120

Warren, D. E., Tranel, D. & Duff, M. C. (2016). Impaired acquisition of new words after left temporal lobectomy despite normal fast-mapping behavior. *Neuropsychologia*, *80*, 165–175. doi: 10.1016/j.neuropsychologia.2015.11.016

Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A. & Gaskell, M. G. (2017). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and Language*, *167*, 13–27. doi: 10.1016/j.bandl.2016.07.010

Welch, B. L. (1947). The generalisation of 'student's' problem when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. doi: 10.1093/biomet/34.1-2.28

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* New York: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org

Wilson, M. (1988). MRC psycholinguistic database: machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, and Computers*, *20*(1), 6–10. doi: 10.3758/BF03202594

Winocur, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioural Brain Research*, *38*(2), 145–154. doi: 10.1016/0166-4328(90)90012-4

Zaiser, A.-K., Meyer, P. & Bader, R. (2019a). Evidence for fast mapping in adults — moderating factors yet need to be identified. *Cognitive Neuroscience*, *10*(4), 232–233. doi: 10.1080/17588928.2019.1605986

Zaiser, A.-K., Meyer, P. & Bader, R. (2019b). *Feature overlap modulates rapid semantic but not lexical integration of novel associations by means of fast mapping.* Retrieved from https://www.biorxiv.org/content/10.1101/594218v1 doi: 10.1101/594218

Zgonnikov, A., Aleni, A., Piiroinen, P. T., O'Hora, D. & di Bernardo, M. (2017). Decision landscapes: visualizing mouse-tracking data. *Royal Society Open Science*, *4*(11), 170482. doi: 10.1098/rsos.170482

Zhang, Q., Popov, V., Koch, G. E., Calloway, R. C. & Coutanche, M. N. (2018). Fast memory integration facilitated by schema consistency. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2777–2782. doi: 10.1101/253393

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. doi: 10.1126/science.2218534

Zosh, J. M., Brinster, M. & Halberda, J. (2013). Optimal contrast: competition between two referents improves word learning. *Applied Developmental Science*, *17*(1), 20–28. doi: 10.1080/10888691.2013.748420