CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Elsevier - Publisher Connector

Cognitive Psychology 79 (2015) 1–39

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

From specific examples to general knowledge in language learning

Jakke Tamminen [a,*], Matthew H. Davis [b], Kathleen Rastle [a]

[a] Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, United Kingdom
[b] Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom

ARTICLE INFO

Article history:
Accepted 23 March 2015
Available online 17 April 2015

Keywords:
Language learning
Generalisation
Morphology
Memory consolidation

ABSTRACT

The extraction of general knowledge from individual episodes is critical if we are to learn new knowledge or abilities. Here we uncover some of the key cognitive mechanisms that characterise this process in the domain of language learning. In five experiments adult participants learned new morphological units embedded in fictitious words created by attaching new affixes (e.g., *-afe*) to familiar word stems (e.g., "*sleepafe* is a participant in a study about the effects of sleep"). Participants' ability to generalise semantic knowledge about the affixes was tested using tasks requiring the comprehension and production of novel words containing a trained affix (e.g., *sailafe*). We manipulated the delay between training and test (Experiment 1), the number of unique exemplars provided for each affix during training (Experiment 2), and the consistency of the form-to-meaning mapping of the affixes (Experiments 3–5). In a task where speeded online language processing is required (semantic priming), generalisation was achieved only after a memory consolidation opportunity following training, and only if the training included a sufficient number of unique exemplars. Semantic inconsistency disrupted speeded generalisation unless consolidation was allowed to operate on one of the two affix-meanings before introducing inconsistencies. In contrast, in tasks that required slow, deliberate reasoning, generalisation could be achieved largely irrespective of the above constraints. These findings point to two different mechanisms of

* Corresponding author.
E-mail addresses: jakke.tamminen@rhul.ac.uk, jakke.tamminen@gmail.com (J. Tamminen), matt.davis@mrc-cbu.cam.ac.uk (M.H. Davis), kathy.rastle@rhul.ac.uk (K. Rastle).

http://dx.doi.org/10.1016/j.cogpsych.2015.03.003
0010-0285/© 2015 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

generalisation that have different cognitive demands and rely on different types of memory representations.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Humans must draw on relevant past experiences in deciding how to respond to a new stimulus or situation. One challenge in this process is that experience occurs as single instances or episodes. We must combine information over multiple instances or episodes to arrive at the more general knowledge that can best guide future behaviour. For example, despite retaining an episodic memory for what the weather was like yesterday and on specific previous days, one's decision about whether to wear a raincoat today may be based on more general knowledge of the prevailing weather conditions given the location, the season, and other factors. Researchers since Tulving (1972) have correspondingly distinguished episodic memories of specific past experiences from semantic memory: a store of abstract general knowledge that is used to guide decision-making, language processing, and other forms of complex behaviour.

This distinction between episodic memory and more abstract, general knowledge is common to many domains of cognition. Accordingly, debate concerning the functional mechanisms responsible for acquiring and expressing these different forms of knowledge has a long history in cognitive science. For example, in learning the structure of novel conceptual categories, two broad classes of theory were initially proposed: exemplar theories in which generalisation is achieved by combining representations of multiple individual instances (Medin & Schaffer, 1978; Nosofsky, 1986) and abstractionist theories in which category representations are structured around more abstract knowledge of the central tendency or prototype derived from many instances (Posner & Keele, 1968). These accounts proved difficult to separate behaviourally (Minda & Smith, 2002; Zaki, Nosofsky, Stanton, & Cohen, 2003; though see Mack, Preston, & Love, 2013, for relevant neural evidence), leading to recent hybrid and multiple mechanism accounts (Kumaran & McClelland, 2012; Love, Medin, & Gureckis, 2004).

Here we use an artificial language learning method to explore the processes that yield general, semantic knowledge from individual learned words. We will begin by describing various forms of generalisation that are apparent in language learning, including the morphological regularities that are the focus of the present work. We will then consider how modern dual-mechanism accounts of episodic and semantic learning accommodate this type of generalisation, before laying out the specific methodology and factors to be explored in the five experiments presented in the paper.

1.1. Generalisation in language and language learning

Language is one domain of complex human behaviour that reflects and requires the acquisition of general knowledge from exposure to individual episodes. These generalisation processes characterise multiple levels of the language system. For example, in respect of single words, we extract information about spelling-to-sound relationships from exposure to existing words (e.g., *moon*, *noon*, *loon*, *soon*), and it is this general knowledge that allows us to decode unfamiliar words and nonwords (e.g., *voon*; e.g., Coltheart, 1978). Similarly, we extract information about how to express past-tense status from exposure to existing words (e.g., *jumped*, *walked*, *watched*), and it is this general knowledge that can lead to over-regularisation errors in young children (e.g., *drinked*, *keeped*, *teached*; e.g., Marcus et al., 1992). In higher levels of language processing, it is our general knowledge of the permissible syntactic structures of language that allows us to understand and to express a near infinite number of phrases, sentences, and ideas (e.g., Tomasello, 2000); and of course, overgeneralisation processes at a conceptual level are at the heart of stereotyping (e.g., Cantor & Mischel, 1979).

One of the most powerful examples of linguistic generalisation at the level of single words arises in the domain of morphology. In English, as in most other languages of the world, we combine stems (e.g., *trust*, *clean*) with a small number of prefixes (e.g., *un-*, *dis-*) and suffixes (e.g., *-er*, *-ly*, *-y*) to form the vast majority of word forms (around 85%; e.g., *distrust*, *trusty*, *cleanly*, *unclean*). The critical property of this kind of combinatorial system is that language users have generalised knowledge of its components: they are able to use the morphemic units of the language flexibly outside of the particular contexts in which they were learned to understand and to produce new meaningful words. For example, the American public had no difficulty understanding George W. Bush when in 2006 he said, "But I'm the *decider*, and I decide what's best", because they knew that the suffix [-er] conveys the meaning "*one who*" in relation to a verb stem, yielding "*one who decides*". Around 70% of new words entering the English language are simple recombinations of existing morphemic units (Algeo, 1991), making morphology key to lexical productivity (e.g., *gamification*, which entered the language in 2012; *hackable* and *tweetable*, which both entered the language in 2013; Oxford English Dictionary, 2014). There is also a broad consensus that our general knowledge about morphemes is central to the recognition (e.g., Marslen-Wilson, Tyler, Waksler, & Older, 1994; Rastle & Davis, 2008) and production (e.g., Treiman & Cassar, 1996; Zwitserlood, Bölte, & Dohmes, 2000) of spoken and written words.

The work presented in this article investigates how it is that we acquire knowledge about affix morphemes that can be generalised outside of the particular contexts in which they were learned. Most broadly, we are interested in the relationship between the learning of individual exemplars (i.e. of words containing a particular affix) and the development of general morphemic knowledge that can be deployed in the service of understanding the meaning of new morphemic constructions (e.g., *decider*, *hackable*). Many models of lexical processing postulate abstract local morphemic representations accessed during word recognition (e.g., Marslen-Wilson et al., 1994; Taft, 1994), perhaps in parallel with whole-word representations (e.g., Caramazza, Laudanna, & Romani, 1988; Schreuder & Baayen, 1997). Other models do not have explicit morphemic representations, yet represent morphological knowledge as 'islands of regularity' in abstract internal units that mediate the mapping between word forms and their meanings (e.g., Plaut & Gonnerman, 2000). Our work seeks to discover more about the processes that give rise to these morphemic representations. Does the development of general morpheme knowledge reflect the simple accumulation of encounters with whole words that contain the relevant morphemes, or are there other more complex constraints on the acquisition of this knowledge?

In order to investigate this problem, we present a laboratory model of morpheme learning in which adults are trained on sets of novel words comprising a familiar stem and a novel affix (e.g., *buildafe*, *sleepafe*, *teachafe*) and are then assessed in a variety of ways as to their knowledge of the novel affix (*-afe*, in this case; see also Merkx, Rastle, & Davis, 2011; Tamminen, Davis, Merkx, & Rastle, 2012). Critically, this method allows us to measure both participants' knowledge of the learning episodes (i.e. their recollection of the individual trained novel words like *buildafe*, *sleepafe*), and the development of generalised affix knowledge that permits understanding of untrained morphemic constructions (e.g., *sailafe*). This method therefore allows us to assess the role of item-specific knowledge of newly-learned words in learners' ability to extract elements of underlying structure that support generalisation to untrained words.

1.2. Generalisation through complementary learning systems

Dual-mechanism theories provide a potentially useful framework in which to consider generalisation processes of learning and memory. As we introduced at the outset, early theories of generalisation relied either on mechanisms that operated solely on episodic memory traces (e.g., Hintzman, 1986, 1988) or on mechanisms that involved the creation of abstract representations during learning (e.g., Posner & Keele, 1968). The more recent dual-mechanism accounts on the other hand postulate two dissociable neural mechanisms for encoding memories of new episodes and for laying down long-term abstract representations of general semantic knowledge (Alvarez & Squire, 1994; Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995; Winocur & Moscovitch, 2011). One particularly influential theory within this class is the Complementary Learning Systems (CLS) account (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Norman, 2002), which offers a detailed neuroanatomical and computational description of these operations. Like other dual-mechanism theories,

this account proposes two learning systems, a hippocampal system that allows fast encoding of episodic memories, and a neocortical system that slowly integrates new memories with existing knowledge for permanent storage.

One key property of the neocortical system is that it generates overlapping representations and uses a single set of weighted connections to store long-term knowledge. In this account, information is represented by the brain (or by a computational model simulating cognitive function) in patterns of activity over a population of neurons (or units in a computational model). Activating semantic representations of two similar concepts (e.g., two birds such as a robin and a canary) thus elicits two similar patterns of activity that share a large number of active neurons. In this sense, then, the two semantically related concepts have overlapping representations. Such overlap allows the neocortical system to represent similarities across memories or large bodies of knowledge. Long-term knowledge in the neocortical system is stored in connection weights (i.e. the strength of the individual connections) that link sets of neurons representing different forms of associated information. Connection weights that support knowledge of one concept (e.g. linking the appearance of a bird to its ability to fly) will function similarly for a set of related concepts. Thus, overlapping neural representations, combined with knowledge stored in overlapping sets of connections provide a powerful means for generalisation of new memories: it allows the system to benefit from shared structure across a number of different exemplars of a given semantic category, and to generalise these properties to novel exemplars of the same category. For example, if a learner is introduced to a new type of bird, s/he is able to generalise typical properties of other, similar birds to the new exemplar (e.g., it is likely to have wings and feathers, it is likely to be able to fly, etc., see Rogers & McClelland, 2004).

Overlapping representations and shared connection weights may come at a significant cost, however. Acquiring and integrating new information with existing information in a system like this must proceed in a gradual manner to avoid interference between new and old knowledge. This means that new information must be added to the neocortical store slowly, through repeated presentations that interleave new learning with exposure to existing knowledge in the same semantic domain. Slow and interleaved learning is necessary for connection weights to be jointly determined by the shared structure of the whole stimulus set rather than by the idiosyncratic properties of the new stimulus alone. Interleaving is important, also, to allow existing information to be retained and also being modified by relevant newly learned information. For example, if one learns additional facts about the vocal organ or "syrinx" of a canary this leads to modifications to connection weights to ensure that this new knowledge can be applied to other familiar birds whilst retaining relevant information on whether these other birds can sing or not.

One illustration of the importance of interleaved learning is that computational models that employ overlapping representations can be susceptible to catastrophic interference (French, 1999; McCloskey & Cohen, 1989; Ratcliff, 1990). In the absence of interleaved learning, the introduction of new information in newly learned items can overwrite pre-existing information and lead to catastrophic forgetting. Interleaving of new and old items, and slow learning can overcome this problem (McClelland et al., 1995). These limitations are highly problematic however, because learning in natural environments typically needs to occur very rapidly on the basis of limited exposure. The CLS account offers a solution to this problem in the form of the hippocampal system which employs pattern separation to generate decorrelated, distinct representations of specific items or episodes. Such distinct representations allow rapid encoding of new information without risk of catastrophic interference between new and old knowledge (see Bowers, Vankov, Damian, & Davis, 2014, for a discussion of how distinct representations solve a related problem of superposition interference in short-term memory). These distinct episodic representations are then used to support the gradual development of overlapping cortical representations during offline periods such as sleep. Thus offline consolidation of hippocampal representations into the neocortex provides a mechanism for avoiding massed exposure to new instances that might otherwise lead to catastrophic interference.

1.3. Is consolidation needed for generalisation in language learning?

Dual-mechanism theories such as the CLS have had some success in helping us to understand how new words become integrated into the mental lexicon (e.g., Davis & Gaskell, 2009; Gaskell & Dumay,

2003) but their application to the problem of generalisation has been more limited. Tamminen et al. (2012) argued that if the CLS account of generalisation is correct, then we should find that a period of consolidation facilitates the emergence of generalisation. They reasoned that individual learning episodes should be encoded rapidly in the hippocampus, but because these representations are non-overlapping, they may not permit all forms of generalisation. If it is the discovery of shared structure across multiple learning episodes that is required for generalisation then this would be supported by the development of shared representations for similar items in the neocortex. According to the CLS theory this process takes time since it depends on either the slow accumulation of new knowledge during learning (with a training protocol that interleaves all relevant new and existing knowledge, cf. Lindsay & Gaskell, 2013), or a form of interleaved learning that can only be achieved during offline consolidation.

Tamminen et al. (2012) sought to test this prediction in the context of the acquisition of morphological knowledge, using the morpheme learning paradigm developed by Merkx et al. (2011). They trained participants on novel words comprising familiar stems and novel affixes (e.g., *sleepafe*). These novel words were associated with semantically-transparent and consistent meanings during training (i.e. the meaning of the novel word was related to the stem, and all instances of *-afe* referred to a person; e.g., "*sleepafe* is a participant in a study about the effects of sleep"). Results provided some support for the CLS prediction: in a test of speeded auditory repetition (shadowing), participants responded faster to an untrained stimulus that contained a trained affix (e.g., *sailafe*) than an untrained stimulus without one (e.g., *sailnept*), but only in a group tested two days after training. Those participants tested immediately after training showed no benefit of having learned the novel affix. This result seems to suggest that consolidation is required for the generalisation of learned morphemic knowledge. However, Tamminen et al. (2012) also conducted a second, non-speeded test of generalisation in which participants were asked to choose between two possible definitions for an untrained stimulus with a trained affix (e.g., participants might have to decide whether *sailafe* could refer to "someone who goes to sea every weekend to sail" or to "the hourly cost of learning how to sail a yacht". If *-afe* in training always occurred in novel words referring to people, participants should indicate the former option to be the more plausible one). These definitions either respected the meaning of the affix in the trained set or took the meaning of a different novel affix. Participants were highly accurate on this task and showed no performance differences as a function of day of testing.

On the basis of these data, Tamminen et al. (2012) argued that consolidation is required for the development of neocortical representations that support generalisation in speeded tasks, while some forms of generalisation in non-speeded tasks can be accomplished on the basis of episodic hippocampal representations (e.g., by combining multiple episodes at the time of testing). However, the two generalisation tests reported by Tamminen et al. (2012) differed not only in their task demands but also in the nature of the information being probed. The test that appeared sensitive to consolidation required a speeded response and probed participants' knowledge of the phonological form of learned affixes. The test that appeared insensitive to consolidation was not performed under time pressure and probed participants' knowledge of affix meaning. Thus, it remains unclear whether consolidation is needed for generalisation, and if so, which aspects of generalisation may require consolidation. Is it that consolidation is required for the development of sufficiently stable representations to impact on speeded lexical processing? Or is it that the generalisation of form but not meaning requires consolidation?

1.4. Further constraints on the generalisation of morphemic information

In the series of experiments reported here, we ask what it takes to acquire knowledge of a novel affix that can be generalised to new (stem) contexts, under speeded and non-speeded conditions. In contrast to our previous investigations of this problem which focused on the acquisition of knowledge about affix *forms* (Merkx et al., 2011; Tamminen et al., 2012), we are particularly interested in the acquisition of *semantic* properties of affixes, which are key to understanding and expressing new words for existing or new concepts. In line with Tamminen et al. (2012), we consider affix learning in the context of CLS theories of memory (e.g., McClelland et al., 1995). As described in detail above, these theories distinguish two learning systems based on their anatomical organisation (hippocampal

vs. neocortical), the representations of individual items (distinct representations for similar items in the hippocampus, overlapping or shared representations for key aspects of similar items in the neocortex) and the time-course of acquisition (rapid learning in the hippocampus, slower learning in the neocortex). However, for the present purposes we focus on a functional distinction that we believe is critical for the learning and generalisation of linguistic knowledge. We follow Kumaran and McClelland (2012) in proposing that generalisation from distinct hippocampal representations requires recurrent activation of multiple item representations and hence additional processing time. In contrast, generalisation from neocortical representations can be achieved more rapidly since distributed representations can similarly process both familiar and novel items (as in parallel distributed processing models of morphological processing (e.g., Rumelhart & McClelland, 1986)). Based on these considerations, then, we predicted that one signature of hippocampal and neocortical language knowledge would be the relative success and failure of generalisation in non-speeded vs. speeded test tasks. To assess this prediction, we therefore designed a paradigm that examines the generalisation of affix meanings, both in a speeded lexical processing situation and in a non-speeded situation.

Experiment 1 investigates the question of whether consolidation is required for the generalisation of affix meanings, and whether its impact is equivalent across speeded and non-speeded tasks by training participants on new affixes and then testing them either immediately after training (i.e. with no consolidation opportunity) or one week after training (thus providing a long consolidation opportunity). Experiments 2–5 go on to examine two factors that are known to be important in adult morphological processing: *contextual diversity* and *semantic consistency*.

Contextual diversity is an important variable for two reasons. Firstly, affixes that combine with many stems (i.e. affixes that benefit from high contextual diversity) are recognised by adult readers more easily than those that combine with few stems (De Jong, Schreuder, & Baayen, 2000; Ford, Davis, & Marslen-Wilson, 2010; Marslen-Wilson, Ford, Older, & Zhou, 1996). Therefore this variable may have a critical impact on learning as well. Secondly, models such as the CLS that use overlapping representations make clear predictions about the impact of contextual diversity on learning, and generalisation in particular. In a system where information is represented as a pattern of activity over a population of processing units (such as neurons in the neocortex) we would expect the semantic information that is shared across several related novel words to gradually become more distinct (or independent of the individual words) as the number of learned words increases. Applied directly to the affix learning paradigm, we expect that as we increase the number of unique stems with which a given novel affix combines in a semantically consistent manner, the pattern of activity associated with the affix becomes increasingly independent of the stems used in training, and consequently increasingly generalizable. In Experiment 2 we test these predictions by training participants on novel affixes that occur with many or few stems in the training session. Critically, unlike in natural language, we can equate the frequency of novel affixes that differ in contextual diversity by manipulating the frequency of words in the training set.

Semantic consistency also impacts on the ease of word recognition and word learning (e.g., Rodd, Gaskell, & Marslen-Wilson, 2002; Rodd et al., 2012) whereby semantically consistent words (words with few unrelated meanings) enjoy a processing benefit. This variable may also affect the overlapping representations that the CLS account requires in order to generalise. Affixes that are associated with multiple different meanings may provide a weaker degree of representational overlap than affixes that consistently refer to the same semantic category. We test this prediction in Experiment 3. In Experiments 4 and 5 we test further predictions of the CLS account about the role of consolidation in allowing acquisition of semantically inconsistent affixes. In sum, the CLS account predicts that any factor which enhances the discovery of this shared structure should benefit the generalisation process, while any factor that disrupts discovery of shared structure should disrupt the generalisation process. We test this general hypothesis by taking advantage of the well-documented effects of memory consolidation, contextual diversity, and semantic consistency on language processing and learning.

1.5. The current experiments

In all the experiments reported here, we trained participants on novel words comprising a familiar stem and a novel affix (e.g., *sleepafe*). Each novel word was given a semantically-transparent definition

relating back to its familiar stem (e.g., "*sleepafe* is a participant in a study about the effects of sleep"). Participants were trained on several different exemplars using each affix, with each affix modifying the meaning of its stem in a semantically-consistent manner (except in those experiments in which we varied semantic consistency). In all cases, however, the meaning of the stem was transparently preserved in the derived form (i.e. we did not include items analogous to *"department"* in which the whole form meaning is not clearly related to the meaning of the stem, cf. Marslen-Wilson et al., 1994). Test tasks were designed to assess both episodic memory for the novel words and general knowledge of the meaning of the novel affixes. To assess participants' episodic memory for the novel words, we employed a yes/no recognition memory task, sometimes in combination with free recall. To assess whether participants had acquired general knowledge of affix meanings that can be accessed in a speeded task, we developed a sentence congruency task, in which participants are asked to read sentence frames that are followed by a semantically-congruent or semantically-incongruent final word. Previous research has demonstrated that semantic congruency between a sentence frame and the final word of a sentence modulates processing time of the sentence-final word, with incongruent words being more difficult to recognise (e.g., Tulving & Gold, 1963) or read aloud (e.g., West & Stanovich, 1978) than congruent words.

Our version of this task involved the speeded reading aloud of sentence-final *untrained* novel words comprising a known stem and newly-learned affix (e.g., *sailafe*). These words were preceded by a sentence frame (e.g., "The manager often argued with the…"), which was semantically-congruent or semantically-incongruent with the meaning of the *affix* (sentence frames were neutral to the meanings of the stems). These sentence-final targets comprised untrained stems with trained novel affixes, so offered an opportunity to assess the extent to which knowledge of the novel affixes generalised to new stems. We took two measures from this task. First, we measured the time taken to read aloud the sentence-final target, which appeared once participants indicated that they had read the sentence frame. Note that this aspect of the task does not *require* generalisation of the affix meanings; the task only requires participants to read the sentence-final pseudoword aloud. However, we reasoned that if participants had developed a stored representation of the novel affix, then their processing of the sentence-final target would be speeded in a semantically-congruent context. Second, we required participants to make a non-speeded explicit judgment about the semantic congruency of the sentence (e.g., "The manager often argued with the *sailafe*" is congruent, if participants have learned that *-afe* refers to a person, but incongruent if they have learned that *-afe* refers to a location, for example). This aspect of the task thus *required* participants to generalise the meaning of the affix. Thus, this task provided two measures of affix meaning generalisation – one measure in a speeded situation in which generalisation of affix meanings was not required (i.e. the congruency effect on reading aloud) and one measure in a non-speeded situation in which generalisation of affix meanings was required (i.e. the accuracy of sentence congruency judgements).

2. Experiment 1

Experiment 1 was designed to address (a) whether exposure to a relatively small set of exemplar words would result in the emergence of generalised knowledge about the meanings of the novel affixes embedded in those words; and (b) the extent to which this generalisation process requires consolidation to be observed in speeded and non-speeded tasks. In order to investigate these issues, we trained our participants on a set of novel affixes combined with familiar stems, and then tested them either immediately after training or one week after training. Participants were tested using a recognition memory task and the sentence congruency task described above.

Using a similar learning paradigm as that used in this study, Merkx et al. (2011) and Tamminen et al. (2012) already established that immediately after training, participants are able to extract the meanings of novel affixes and apply them to new exemplars in a non-speeded task that requires explicit generalisation. Based on these findings, we would expect participants in our study to be able to perform the non-speeded explicit congruency judgement task immediately following training. The more intriguing question concerns whether a congruency effect will be observed on reading aloud latencies, and whether this effect will require consolidation to emerge. While Tamminen et al. (2012) claimed

that consolidation is required for generalisation to emerge in speeded language processing tasks, their evidence was based on a task that does not test access to semantic representations (single-word speeded shadowing). Critically, accounts of generalisation implementing dual-mechanisms predict that any semantic congruency effect on reading aloud latencies should not be observed immediately after training, because the discovery of shared semantic structure requires memory consolidation in order for newly learned information to develop overlapping neocortical representations.

2.1. Method

2.1.1. Participants

Forty-eight native English-speaking participants were recruited. None reported suffering from disorders affecting reading or hearing. Twenty-four were tested immediately after training (15 female, 3 left-handed, mean age = 20), and 24 were tested one week after training (18 female, 1 left-handed, mean age = 21). All were students at Royal Holloway, University of London, and were paid for their participation.

2.1.2. Materials

2.1.2.1. Learning phase. Eight novel affixes were selected from the set used by Merkx et al. (2011). These fell into four orthographic structure groups: CVCV (*nule, tege*), VCV (*afe, ude*), CVCC (*lomb, halk*), and VCC (*esh, ort*). The set was divided into two lists of four affixes each, with all CV structure types used in both lists. One list was trained and the other remained untrained. The assignment of these lists to training condition was counterbalanced across participants.

Thirty-two existing words were selected as stems for the trained novel words (yielding 8 per trained affix). All stems were monosyllabic, monomorphemic, 3–5 letters long ($M = 4.13$), and of sufficiently high CELEX frequency (Baayen, Piepenbrock, & van Rijn, 1993) to be familiar to all speakers of British English ($M = 22.92$ occurrences/million). Audio recordings were made of each trained novel word by a female native speaker of southern British English.

Thirty-two definitions were created for the trained novel words. Definitions were constructed by using the meaning of each novel affix consistently to modify the meaning of each stem. The trained novel affixes were each assigned to one meaning category based on the meanings captured by existing affixes (see also Merkx et al., 2011; Tamminen et al., 2012). These included a place (e.g., -ery in *bakery, nunnery*), a tool (e.g., -er in *cooker, eraser*), a cost (e.g., -age in *postage, corkage*), and a person (e.g., -ist in *creationist, feminist*). Thus, meanings of the novel words were always transparently related to the meanings of their stems, and each affix was used in a semantically-consistent manner across the training set. Sample definitions are provided in Table 1.

2.1.2.2. Testing phase. Participants' knowledge of the novel affixes and novel words was assessed using a sentence congruency task and a recognition memory task.

In the sentence congruency task, participants read aloud 64 novel word targets preceded by a sentence frame (5–11 words long) that was semantically congruent or incongruent with the meaning of the novel affix. The target novel words for reading aloud were formed by combining an *untrained* existing word stem (not seen during the learning phase) with a *trained* novel affix. This formation allowed us to make inferences about participants' generalised knowledge of the novel affixes. The novel words for reading aloud were divided into two lists, one to be used in the congruent condition and the other to be used in the incongruent condition, with lists rotated through these conditions across participants. The stems in the two lists were closely matched for frequency (77.59 vs. 79.74), orthographic neighbourhood size (7.00 vs. 6.91), number of letters (4.03 vs. 3.97), and average production latency of the initial phoneme (270.40 ms vs. 270.33 ms; Rastle, Croot, Harrington, & Coltheart, 2005). Each of the sentence frames was congruent with one of the four meanings for the novel affixes, and was incongruent with the other three meanings. Sentence frames were all semantically neutral in respect of the *stem* of each novel word target. Sentence frames in the incongruent condition were created by scrambling congruent sentence + novel affix pairings, thus creating a semantic mismatch between the sentence frame and the novel affix. In this way, the same stimuli (sentence frames and

Table 1
Examples of trained affixes and stems, their associated meanings, and untrained affixes in one counterbalancing list.

Affix	Examples of trained novel words and associated meanings
-nule	Bricknule is the labourer who operates the oven which hardens clay to brick Foxnule is someone who looks after a fox harmed in a car accident
-afe	Crabafe is the zoo building where you can see exotic crab species Gunafe is the section of an armoury where one can find a gun
-lomb	Fetchlomb is an extendable arm used to fetch small items without getting up Mowlomb is a popular machine which can mow the lawn automatically
-esh	Warnesh is the yearly cost the state pays to warn expatriates of danger Begesh is the amount children pay to gangsters to be allowed to beg

Note: The untrained list of affixes in this example consist of -tege, -ude, -halk, -ort.

Table 2
Examples of sentences used in the sentence congruency task.

Congruency condition	Sentence	Meaning of the affix
Congruent	It was honour to be visited by the sandnule	Person
	The man rushed to get inside the beanafe	Place
	They were taught how to operate a warmlomb	Tool
	He thought it would help if he paid them a hurlesh	Cost
Incongruent	The company had just relocated to a peachnule	Person
	The manager often argued with a pigafe	Place
	They were arrested for not paying the required lynchlomb	Tool
	They always made fun of her for using a readesh	Cost

novel words) were used in both congruent and incongruent conditions, counterbalanced across participants (see Table 2 for examples).

The recognition memory task involved presentation of successive novel words, and required participants to indicate whether these novel words were trained in the learning phase or untrained (never encountered before). The stimuli included all 32 trained novel words from the learning phase (e.g., *sleepafe*, *bricknule*) as well as 64 untrained novel words. These 64 untrained novel words were constructed by combining trained and untrained stems and affixes in different configurations. They consisted of 16 untrained stem + trained affix combinations (e.g., *rockafe*), 16 trained stem + untrained affix combinations (e.g., *bricktege*), and 32 recombinant novel words created by combining a trained stem with a trained affix which during the learning phase occurred with a different stem (recombinant stem + affix combinations, e.g., *brickafe*, *sleepnule*). Therefore this task required selection of further 16 monosyllabic word stems which did not occur in the learning phase (mean frequency = 46.10, mean number of letters = 4.13).

2.1.3. Procedure

Participants completed the learning phase before moving on to the testing phase, which was conducted either immediately following learning or after a one week delay. Learning and testing tasks were carried out on computers running DMDX (Forster & Forster, 2003), with standard keyboards used for response collection in the learning phase, and button boxes in the test phase.

2.1.3.1. Learning phase. It is now well known that long-term memory benefits from retrieval practice during a study session (e.g., Karpicke & Roediger, 2008). Therefore, to optimise the effectiveness of our training regime, our learning phase included two different tasks: a typing task involving only study of the materials and an active recall task involving retrieval practice. In the typing task, a novel word and its definition were presented on the screen. The novel word was simultaneously heard once through headphones. Participants were instructed to read the word and its definition, and then press a key. The

key press cleared the screen, and participants then typed the novel word. No time limit was imposed in this task. In the recall task, a novel word definition was presented on the screen, and participants were required to recall and type the novel word that corresponded to that definition. No time limit was imposed. After this response, the correct novel word was displayed, allowing participants to use each recall trial as a learning opportunity, irrespective of whether they could recall the word at that time. The learning phase consisted of nine blocks of the typing task, with each novel word seen once in each block. These nine blocks were interleaved by three recall blocks, always presented after a sequence of three typing blocks. Each novel word appeared once in each recall block. Therefore, each of the 32 novel words was encountered in total 12 times in the learning phase. This phase lasted roughly one hour.

2.1.3.2. Testing phase. In the test phase, participants completed the sentence congruency task and then the recognition memory task. The sentence congruency task started with the visual presentation of a sentence frame with the final word missing. Participants were asked to read the sentence frame silently and to press a button on the button box once they had finished. Upon the button press, the screen was cleared and the final word of the sentence presented in the middle of the screen. The task was now to read the final word aloud as quickly and as accurately as possible. The vocal response was recorded by DMDX through a head mounted microphone. Four seconds after the onset of the word it was replaced by the question "Did the sentence make sense?", upon which participants pressed a button labelled "Yes" or "No" on the button box. In the experimental trials, the final word of the sentence was always an untrained novel word, but these were preceded by a block of four practice trials where familiar words were used instead to familiarise participants with the task. Order of the experimental trials was newly randomised for each participant by DMDX. This task took about 10 min to complete.

In the recognition memory task, a trial started with a fixation cross in the middle of the screen for 500 ms. This was followed by the presentation of one of the trained or untrained novel words. Participants indicated whether the word was trained or untrained on the button box. They were asked to respond as quickly and as accurately as possible within an eight second response window. Order of trials was randomised, and the task lasted about five minutes.

2.2. Results

Reaction time (RT) data were analysed using mixed-effects modelling (Baayen, Davidson, & Bates, 2008) in R using the *lme4* package. This decision allowed us to include participants and items simultaneously in the same model. Random effects structure was always determined by comparing a series of models with gradually simplifying structure, thus preserving those factors that contributed significantly to the model fit. In *lme4* there is at the time of writing no way to calculate *p*-values to evaluate significance for mixed-effect models that include random slopes. We adopted the solution of carrying out likelihood ratio tests to evaluate the significance of each fixed effect by comparing a model which includes the effect to an identical model which does not include the effect (Barr, Levy, Scheepers, & Tily, 2013). Accuracy data were analysed according to the same strategy using logistic mixed-effects models. Here, the *p*-values are reported based on the Wald *Z* statistic for each fixed effect (Jaeger, 2008). Finally, fixed effects were always centred to make main effects interpretable in models including interactions.

2.2.1. Training

To ensure that both time-of-testing groups learned the novel words equally well, we analysed performance at the end of training, in the last block of the recall task. The immediately tested group recalled 96% (±0.99%[1]) of the words, while the delayed test group recalled 97% (±0.70%). A logistic mixed-effects model with no random slopes showed no difference between accuracy in the two groups ($p = .30$).

[1] When we report means in the text, figures, or tables, these are accompanied by the standard error of the mean (indicated by error bars in figures and the ±sign in text and tables). When the means are compared in a within-participants design, we report the within-participant standard error which removes variance due to individual differences (O'Brien & Cousineau, 2014).

2.2.2. Sentence congruency

Reading aloud latencies to the sentence-final words were manually determined using CheckVocal (Protopapas, 2007). Erroneous responses (mispronunciations, hesitations, false starts; 5.3% of the data in the immediately tested group, 5.9% in the delayed test group) were removed, as were extremely long RTs (above 1500 ms; 2.7% of the data in the immediately tested group, 2.3% in the delayed test group). The data were then log-transformed to better meet the assumption of normality and to reduce the effect of remaining outliers. Data in all figures have been backtransformed to show the numbers in an interpretable form. Congruency (congruent vs. incongruent) and time-of-testing (immediate vs. delayed) were included as fixed factors. No random slopes were included, as they did not significantly improve the model fit. The factor of congruency contributed significantly to the model, $\chi^2(1) = 10.76$, $p = .001$, but time-of-testing did not, $\chi^2(1) = 0.09$, $p = .77$. Importantly, the interaction between the two was significant, $\chi^2(1) = 4.89$, $p = .03$. This interaction reflected the fact that while there was no congruency effect observed for the group tested immediately after training, $\chi^2(1) = 0.56$, $p = .45$, there was a significant congruency effect for the group tested a week after training, $\chi^2(1) = 15.14$, $p < .001$. This interaction is depicted in Fig. 1A.

For the analyses of accuracy scores in the congruency judgement component of the sentence congruency task, two participants' data were excluded on the basis of very high error rates (one in the delayed test condition for responding "does not make sense" on every trial, and one in the immediate test condition for responding "does not make sense" in more than 99% of the trials). This pattern of responding suggests that these participants did not understand the task correctly. Accuracy was very high in both the immediately tested group (87 ± 2.10%) and the delayed test group (88 ± 2.25%). Time-of-testing was included as a fixed factor, no random slopes were included. Accuracy in both conditions was reliably above chance (both *ps* < .001), and no difference was observed between the time-of-testing groups, $z = 0.15$, $p = .88$.

2.2.3. Recognition memory

Following Merkx et al. (2011) and Tamminen et al. (2012), we analysed recognition memory data by calculating signal detection measures (*d′*) in order to take into account response bias. Memory of novel word stems was evaluated by calculating the difference between *z*-transformed proportion of

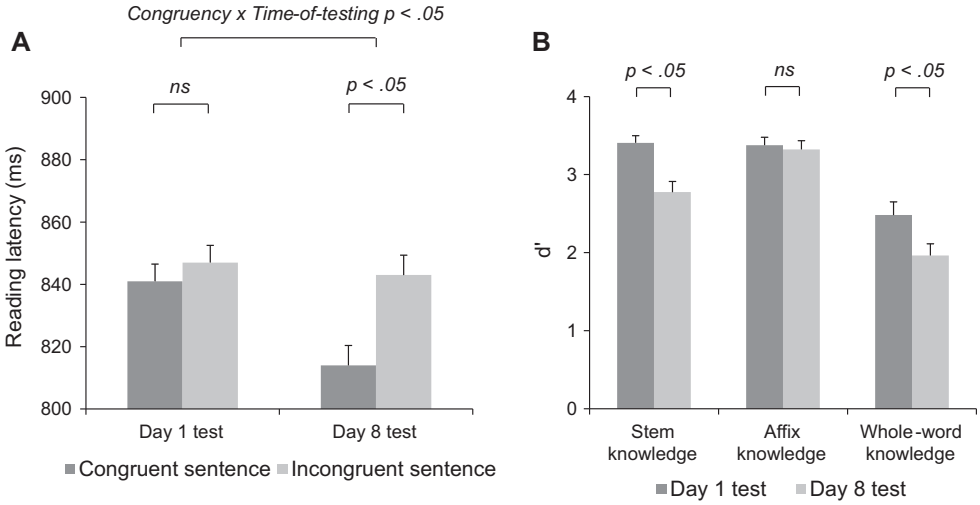

Fig. 1. (A) Reading latencies in Experiment 1 to words in semantically congruent or incongruent sentence contexts for participant groups tested immediately after training (day 1) and one week after training (day 8). Error bars represent within-participant standard error of the mean. (B) *d*-prime scores for the different knowledge types in the recognition memory task for participants tested immediately after training and a week after training. Error bars represent between-participant standard error of the mean. ns = not significant.

accurate "yes" responses to trained novel words (hits) and incorrect "yes" responses to items with an untrained stem and a trained affix (false alarms). Memory of newly learned affixes was measured by calculating the difference between hits and incorrect "yes" responses to items with a trained stem and an untrained affix. Memory for trained stem and affix pairings (i.e. whole word knowledge) was calculated by comparing hits with "yes" responses to recombinant items. These data are presented in Fig. 1B. Since item-level data are not available when analysing d' values, we used analyses of variance (ANOVAs) by participants. An ANOVA with knowledge type (stem, affix, whole-word) as a within-participants factor and time-of-testing (immediate vs. delayed) as between-participants factor showed a main effect of knowledge type, $F(2,92) = 147.60$, $p < .001$, a main effect of time-of-testing, $F(1,46) = 5.81$, $p = .02$, and an interaction between the two factors, $F(2,92) = 9.90$, $p < .001$. To unpack this interaction, we first assessed the effect of time-of-testing in the three knowledge type conditions. This showed that stem knowledge declined significantly over time, $t(46) = 3.85$, $p < .001$, as did whole word knowledge, $t(46) = 2.31$, $p = .03$. Affix knowledge did not change over time. Next we evaluated the difference between knowledge types in both time-of-testing conditions. In the immediate test condition whole-word knowledge was significantly poorer than either stem, $t(23) = 8.19$, $p < .001$, or affix knowledge, $t(23) = 7.44$, $p < .001$. This was also true in the delayed test condition, $t(23) = 9.20$, $p < .001$ and $t(23) = 14.45$, $p < .001$ respectively. Here also the difference between affix and stem knowledge was significant, $t(23) = -5.18$, $p < .001$, with affix knowledge being better than stem knowledge. Accuracy rates (in percent correct) in this task are presented in Appendix A (for all subsequent experiments as well).

2.3. Discussion

The results of Experiment 1 provided two sources of evidence that participants were able to generalise meanings of the novel affixes that they had learned to new stems. First, consistent with our observations from similar offline generalisation tasks in previous work (Merkx et al., 2011; Tamminen et al., 2012), participants were able to judge the sentential congruency of words formed with novel affixes with a high degree of accuracy. Second, participants revealed a congruency effect on reading aloud latencies for the sentence-final target words. Target words in which the novel affix was semantically congruent with the sentence frame were read aloud more quickly than target words in which the novel affix was semantically incongruent with the sentence frame. Recall that the sentences were congruent or incongruent with regard to the meaning of the affix rather than the stem of the sentence-final target word; therefore the congruency effects reflect participants' generalised semantic knowledge of the affix.

Interestingly, the time course of these two indices of generalisation differed. Participants were highly successful in making explicit congruency decisions immediately after training, and performance was statistically equivalent in the group tested one week after training. However, the semantic congruency effect in reading aloud was only observed in the group of participants who were tested a week after training. This result suggests that the affix representations necessary to permit generalisation in this task are not available immediately after training; instead, it appears that the acquisition of these representations requires a period of memory consolidation. This dual time-course was also seen in the generalisation data reported by Tamminen et al. (2012) using speeded shadowing. In both cases, participants were highly accurate at generalising immediately after training in a non-speeded, explicit meaning judgement test, but a period of consolidation was required in order to observe generalisation effects in a speeded test that did not explicitly require participants to generalise. The present findings add to Tamminen et al. (2012) by showing that this difference is not due to test tasks that tap different types of knowledge (form-based vs. meaning-based), but rather due to a difference in the mechanisms that support generalisation before and after consolidation.

While the passing of time between training and testing appears to be beneficial for the emergence of generalisation in affix learning, the recognition memory data showed a detrimental effect of time for episodic memory. Recognition accuracy for the stems used in training (stem knowledge) as well as memory of which stems occurred with which affixes (whole word knowledge) declined over the week intervening between training and testing (Fig. 1B).

It is also of interest to note that we observe a clear difference between the two components of the sentence congruency task: the ability to generalise in the non-speeded explicit congruency judgement component, and the reading latencies to the sentence-final words. While the reading aloud component showed no effect of congruency immediately after training (and therefore no evidence for generalisation), participants were able to generalise in the same session in the explicit judgement task. This is unlikely to have been caused by reading aloud being a less sensitive task to measure generalisation; after all, this task does successfully demonstrate generalisation in the delayed test condition. Instead, we argue that this result suggests that generalisation can be achieved by relying on at least two different mechanisms, depending on the demands of the particular task. We will return to this issue, and its implications for existing theories of generalisation, in Section 7.

3. Experiment 2

As discussed in Section 1, effects of contextual diversity (i.e. the number of contexts in which a sound sequence or letter string arises) have been observed at multiple levels of the language system. Contextual diversity facilitates the recognition of printed words (Adelman, Brown, & Quesada, 2006; McDonald & Shillcock, 2001); it enhances rule extraction in infants and adults (under the terminology 'relative frequency'; Valian & Coulson, 1988 and 'variability'; Gómez, 2002); and it plays a powerful role in morphological processing (under the terminology 'family size'; De Jong et al., 2000). In respect of this latter body of evidence, it is well known that the recognition of morphologically-simple and morphologically-complex words is faster when the stem is from a large morphological family (i.e. is embedded in many morphologically-complex words; e.g., *sign*) than when it is from a small morphological family (i.e. is embedded in few morphologically-complex words; e.g., *skull*; e.g., De Jong et al., 2000). Further, there is evidence that these effects extend to affixes, with affixes that combine with many stems being recognised more easily than those that combine with few stems (Ford et al., 2010; Marslen-Wilson et al., 1996).

In this experiment we investigate whether the diversity of stems with which an affix is trained influences generalisation of that affix. In addition to shedding light on whether the morphological family size effects observed in adult language processing may arise as a consequence of the morpheme acquisition process, this experiment tests a critical prediction made by CLS and other dual-mechanism accounts of memory. Recall that in these accounts generalisation emerges from stored knowledge of the shared properties of events encoded in memory, made possible by the overlap between similar memories. Family size in semantically consistent affixes (i.e. affixes which carry the same meaning across many different words) constitutes a natural metric of such overlap: affixes with a large family size are associated with a large number of lexical representations which all share a common element (the novel affix). This should allow the system to extract a context-independent representation of the affix which should be readily generalizable. Affixes with a small family size benefit much less from the overlap in the neocortical architecture, and are consequently less likely to form representations that are independent of the stem-contexts in which they were trained. We predict that an affix that participates in many different novel words (i.e. occurs with many stems) will be discovered more easily, and thus generalises to a greater degree than an affix that participates in few novel words (i.e. occurs with few stems). Essentially, each additional context that an affix arises in offers greater opportunity for learning the meaning of that affix.

To test these predictions, we again trained participants on new affixes. We varied family size so that half of the affixes occurred during training with eight different stems (large family size), while the other half occurred with two different stems (small family size). However, the number of exposures to each affix was identical across the two family size conditions. To achieve equal number of exposures to the small-family affixes, novel words in this condition were repeated four times as often as words containing large-family affixes. This manipulation should enhance episodic memory for the novel words and affixes in the small-family condition, allowing us to potentially dissociate episodic memory strength from emerging generalisation effects. Thus, unlike in natural language, we can compare the impact of contextual diversity on affix learning and affix representations whilst matching for affix frequency.

3.1. Method

3.1.1. Participants

Twenty-four native English-speaking participants were recruited (17 female, 5 left-handed, mean age = 21). All were students at Royal Holloway, University of London, and paid for their participation.

3.1.2. Materials

3.1.2.1. Learning phase. Sixteen novel affixes were selected. These fell into the same four orthographic structure groups used in Experiment 1: CVCV (*nule, tege, labe, hoke*), VCV (*ane, ose, ude, ete*), CVCC (*nept, tund, lomb, halk*), and VCC (*ort, aph, esh, uck*). This set was divided into two lists of eight affixes each with all CV structure types used in both lists. Each participant learned the affixes of one list only; the other remained untrained. Allocation of lists to these conditions was counterbalanced across participants. Sixty-four existing words were used as stems in the trained novel words. All were monosyllabic, monomorphemic, 3–5 letters long ($M = 4.30$), and of high enough CELEX frequency (Baayen et al., 1993) to be familiar to speakers of British English ($M = 49.61$ occurrences/million).

Thirty-two of the stems (eight stems for each of the four affixes) were assigned to the large family size condition, and another 32 to the small family size condition (counterbalanced so that all stems were used in both conditions). However, in the small family condition, for any given participant only eight of the stems were used, two stems for each of the four affixes. Across participants, however, all of the stems were used. Audio recordings were made of each trained novel word by a female native speaker of southern British English.

Definitions were created for the novel words using the same four meaning categories as in Experiment 1. Again, each affix was consistently assigned to one meaning category. Two meaning categories were used in the small family size condition, and another two in the large family size condition. These were counterbalanced so that all meaning categories were used in both family size conditions. Because we were training each participant on eight novel affixes (four in each family size condition) but had only four meaning categories, each meaning category was assigned two different affixes (see Table 3 for an illustration).

3.1.2.2. Testing phase. The sentence congruency task was designed in the same way as in Experiment 1, except that the stimulus set was doubled to allow us to test for congruency effects in the two different family size conditions. For this purpose we collected 128 monosyllabic word stems which were divided into two lists matched on frequency (21.15 vs. 20.94 per million; Baayen et al., 1993), orthographic neighbourhood size (6.31 vs. 5.91), number of letters (4.36 vs. 4.27), and the average production latency of the initial phoneme (267.11 ms vs. 267.57 ms; Rastle et al., 2005). These two lists were assigned to the congruent or incongruent conditions. The lists were further divided into two sub-lists, matched on the same variables as above, to be used in the two family size conditions. All lists were rotated through the family size and congruency conditions for a fully counterbalanced design. Sentence frames were 4–12 words long, and semantically neutral to the stem.

Table 3
Examples of novel words used in Experiment 2.

Family size	Novel word	Meaning category
Small	Buildnule	Cost
Small	Bringane	Tool
Small	Crewose	Cost
Small	Girlhalk	Tool
Large	Knitlomb	Person
Large	Creepesh	Place
Large	Hairuck	Person
Large	Sheeptege	Place

Note: The untrained list of affixes in this example consist of -nept, -ort, -labe, -ude, -tund, -aph, -hoke, -ete.

In the recognition memory task participants saw all 40 trained novel words from the learning phase as well as 40 recombinant novel words, created in the same was as in Experiment 1. This allowed a measure of whole word knowledge, but the other types of knowledge measured in Experiment 1 were not included here due to the small number of trained stems in the small-family condition.

3.1.3. Procedure

Participants completed the learning phase followed by the testing phase conducted one week later to ensure that all novel words had been consolidated. The same equipment was used as in Experiment 1.

3.1.3.1. Learning phase. The learning phase was identical to that of Experiment 1 with the same tasks and number of exposures to each novel word. The major difference was in the number of words to be learned. In this experiment participants learned a total of 40 novel words: 32 in the large family size condition (eight stems per affix) and eight in the small family size condition (two stems per affix). Each large-family word was encountered 12 times, as in Experiment 1, once in each of the nine typing task blocks, and once in each of the three recall blocks. This means that each large-family affix was encountered 96 times in total (8 stems $\times$ 12 training blocks = 96 exposures). To match the number of encounters with each large and small-family affix, we presented each small-family novel word four times in each of the learning blocks (2 stems $\times$ 12 training blocks $\times$ 4 repetitions = 96 exposures). This phase lasted roughly 90 min, and the presentation order within each block was randomised. The learning phase ended with a free recall test. Participants were given a blank sheet of paper and asked to write down as many of the newly learned words as possible. Recall of word meanings was not required. Each participant was given five minutes to complete this task.

3.1.3.2. Testing phase. The test phase included sentence congruency and recognition memory tasks. These tasks were conducted in an identical manner to Experiment 1.

3.2. Results

3.2.1. Training

To investigate whether both family size conditions were learned equally well, we analysed performance in the last block of the recall task in the training session. Participants recalled an average of 90% (±0.75%) of the large-family words, and 94% (±0.75%) of the small family words. A logistic mixed-effects model with no random slopes showed this difference to be significant, $z = 3.61$, $p < .001$.

3.2.2. Free recall

At the end of the training session, participants recalled on average 48% (±1.21%) of the large family size words, and 79% (±4.88%) of the small family size words. A logistic mixed effects model (with no random slopes) showed this difference to be significant, $z = 7.47$, $p < .001$.

3.2.3. Sentence congruency

As in Experiment 1, incorrect responses (0.7% of the data) and extremely long RTs (above 1500 ms; 4.3% of the data) were removed. Congruency (congruent vs. incongruent) and family size (small vs. large) were included as fixed factors. No random slopes were included. The factor of congruency was not significant, $\chi^2(1) = 1.72$, $p = .19$, while family size did reach significance, $\chi^2(1) = 7.22$, $p = .007$. Critically, we observed an interaction between these two factors, $\chi^2(1) = 4.09$, $p = .04$. This interaction reflected a significant congruency effect in the large family size condition, $\chi^2(1) = 5.66$, $p = .02$, but no congruency effect in the small family size condition, $\chi^2(1) = 0.20$, $p = .65$. These data are depicted in Fig. 2A.

Accuracy scores in the congruency judgment component of the sentence congruency task were analysed to see if family size affected this more explicit measure. One participant's data were excluded from this analysis for responding "does not make sense" on every trial. Family size was included as a fixed factor; random slopes for the effect of family size significantly benefited the model and were

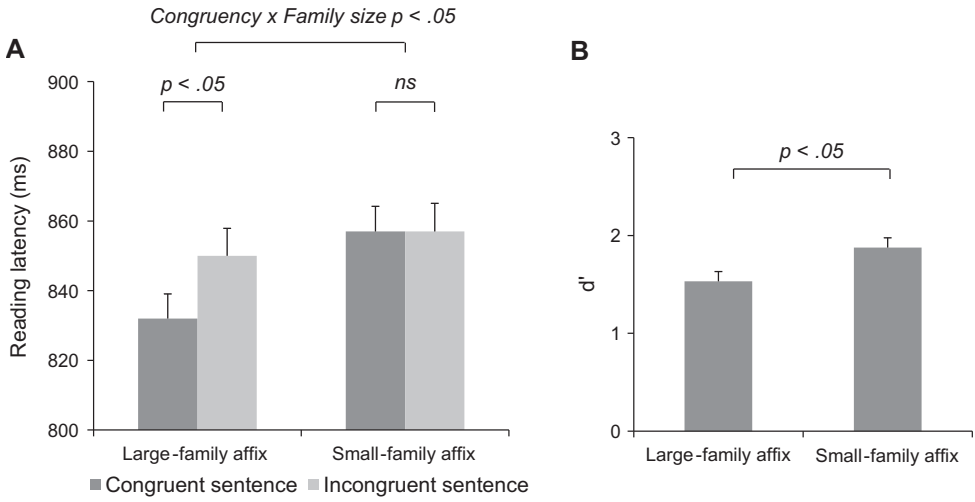

Fig. 2. (A) Reading latencies in Experiment 2 to words in semantically congruent or incongruent sentence contexts for affixes in the two family size conditions. (B) *d*-prime scores for whole word knowledge in the recognition memory task. Error bars represent within-participant standard error of the mean. ns = not significant.

therefore included. Accuracy in both the large family size (70 ± 1.59%) and small family size (67 ± 1.59%) conditions was reliably above chance (both *ps* < .001). A trend-level difference was observed between the conditions, $z = -1.70$, $p = .09$, with slightly higher accuracy in the large-family condition.

3.2.4. Recognition memory

Whole word knowledge was calculated using signal detection measures. A paired-samples *t*-test comparing d' between the large and small family size conditions showed that accuracy was significantly higher in the small family size condition, $t(23) = -2.45$, $p = .02$ (Fig. 2B).

3.3. Discussion

In Experiment 2 we used the sentence congruence task to discover if generalisation for newly learned affixes is modulated by the number of stems with which an affix occurs (critically, while holding affix *frequency* constant). Results again showed different patterns for the two measures of generalisation, much like in Experiment 1. While low contextual diversity did not significantly reduce participants' ability to generalise in the explicit congruency judgment task, it did influence the size of the congruency effect on reading aloud. Specifically, we observed a robust congruency effect for affixes paired with many stems (replicating the delayed test condition of Experiment 1), which was statistically greater than the null effect of congruency observed for affixes paired with few stems.

While generalisation of novel affixes in the speeded reading task required the high degree of variability afforded by multiple stems, episodic memory did not benefit from this. In fact, we observed a clear dissociation between generalisation and strength of episodic memory. Episodic memory benefitted from frequency of presentations during training, as shown by free recall rates being superior in the small-family condition, and the recognition memory task showing significantly better whole-word knowledge for the small-family words than for the large-family words. It therefore appears that accumulation of learning episodes benefits episodic memory measures, but not generalisation, with generalisation critically depending on the number of unique exemplars (unique stem + affix combinations in our case) encountered during training.

4. Experiment 3

Semantic consistency, that is, the consistency with which a word or a morpheme refers to one meaning or semantic class, has a profound impact on language processing and there is emerging evidence to show that it might also influence language acquisition. For example, many words have multiple unrelated meanings (e.g., *bank*); this form-to-meaning ambiguity slows word recognition (e.g., Rodd et al., 2002). It is also more difficult to learn a new meaning for an existing word if that meaning is inconsistent with the existing meaning than if it is consistent with it (for example, learning that *hive* is a busy household is easier than learning that *grin* is a busy household; Rodd et al., 2012). Similarly, form-to-meaning consistency facilitates word learning, with learning being easier when orthographic form predicts the semantic category of a new word (Rueckl & Dror, 1994). In the domain of morphology form-to-meaning consistency of affixes (i.e. whether an affix modifies the stems with which it occurs in a consistent manner) influences the way that morphologically complex words using existing affixes are stored and parsed (e.g., Bertram, Schreuder, & Baayen, 2000). Given that semantic inconsistency abounds at several levels of language, it is critical to understand how it impacts on learning and generalisation, and what the relevant implications are for theories of generalisation. In Experiments 3–5 we seek to establish whether and under what circumstances participants are able to learn affixes with multiple unrelated meanings, and whether generalisation of these affixes occurs in spite of this inconsistency.

Dual-mechanism theories of memory such as the CLS account are faced with a unique challenge during the acquisition of semantically inconsistent information. This challenge arises because in architectures using overlapping representations, the overlapping features of new memories become stronger as more materials with a high degree of similarity are encoded, while individuating features become more weakly represented (Norman & O'Reilly, 2003). This characteristic of such models suggests that semantically inconsistent affixes with multiple different meanings might result in weaker representations than semantically consistent affixes, perhaps disrupting performance in generalisation tasks. This prediction is supported by Plaut and Gonnerman's (2000) connectionist model of morphology which showed little morphological priming for semantically opaque words in a morphologically impoverished language (modelled after English).

In Experiment 3 we compare the acquisition of semantically consistent and inconsistent novel affixes. Participants were trained on novel affixes, half of which referred to one semantic category (e.g., tools) in some novel words and referred to a different semantic category (e.g., people) in other novel words. These semantically inconsistent affixes are common in natural language; for example the affix *-er* can combine with *wild* to form the comparative *wilder*, and can also combine with *teach* to refer to a person *teacher*. The other half of our novel affixes referred to one and the same semantic category in all words (semantically consistent affixes). We predict that semantically consistent affixes should support speeded generalisation following consolidation, as this condition replicates the previous two studies in which all the novel affixes were semantically consistent. However, as described above, generalisation might be disrupted in semantically inconsistent affixes. By comparing the outcome of speeded and non-speeded generalisation tests we can explore whether each of the two mechanisms that potentially support generalisation in dual-mechanism models are differentially affected by semantic inconsistency.

4.1. Method

4.1.1. Participants

Twenty-four native English-speaking participants were recruited (16 female, 2 left-handed, mean age = 20). All were students at Royal Holloway, University of London, and paid for their participation.

4.1.2. Materials

4.1.2.1. Learning phase. The same 16 novel affixes were used as in Experiment 2. This set was divided into two lists of eight affixes with all CV structure types used in both lists. Each participant learned the affixes of one list only, the other remained untrained. These two lists were further divided into two

Table 4
Examples of novel words used in Experiments 3–5.

Semantic consistency	Novel word (meaning category)
Consistent	Buildnule (cost), Sleepnule (cost)
Consistent	Bringane (place), Lockane (place)
Consistent	Crewose (tool), Bombose (tool)
Consistent	Girltege (person), Graintege (person)
Inconsistent	Knitlomb (person), Swimlomb (tool)
Inconsistent	Creepesh (place), Grabesh (cost)
Inconsistent	Hairuck (cost), Gunuck (person)
Inconsistent	Sheephalk (tool), Creamhalk (place)

Note: these are examples from one counterbalancing list. Across the complete group of participants all affixes and stems were used in both consistency conditions.

sub-lists, one used in the semantically consistent, and the other in the inconsistent condition, with two CV structure types in each sub-list. Allocation of all lists to these conditions was counterbalanced across participants such that every affix served in all conditions across the trained/untrained and consistent/inconsistent manipulations.

The same 64 existing word stems were used as in Experiment 2. Thirty-two of the stems (eight stems for each of the four affixes) were assigned to the semantically consistent condition, and another 32 to the semantically inconsistent condition (counterbalanced so that all stems were used in both conditions).

Definitions were created for the novel words using the same four meaning categories as before. Each semantically consistent affix was associated with one meaning category. Each semantically inconsistent affix on the other hand was associated with two different meaning categories. As Table 4 illustrates, each meaning category was associated with three affixes, one consistent and two inconsistent.

4.1.2.2. Testing phase. The sentence congruency task was designed in the same way as in Experiment 2, and used the same stimuli.

Stimuli in the recognition memory task included all 64 trained novel words from the learning phase and 128 untrained novel words. The untrained novel words consisted of 32 untrained stem + trained affix combinations, 32 trained stem + untrained affix combinations, and 64 words created by combining a trained stem with a trained affix which during the learning phase occurred with a different stem (recombinant stem + affix combinations). The untrained stems had an average frequency of 46.49 per million (Baayen et al., 1993) and an average length of 4.88 letters.

4.1.3. Procedure

The learning phase was followed by the testing phase conducted one week later. The same equipment was used as in previous experiments.

4.1.3.1. Learning phase. Participants learned a total of 64 novel words: 32 in the semantically consistent condition (eight stems per affix) and 32 in the semantically inconsistent condition (eight stems per affix). All words were encountered 11 times, once in each of eight typing task blocks, and once in each of three recall blocks. The blocks were ordered as follows: three blocks of typing, one block of recall, three blocks of typing, one block of recall, two blocks of typing, one block of recall. The decision to drop one block of typing (compared to previous experiments) was taken to ensure the training phase in this experiment which involves learning a larger number of words than the preceding experiments would not exceed two hours for even the slower participants. This phase lasted on average 90 min, the presentation order within each block was randomised, therefore meaning that exemplar words representing the two alternative affix-meaning mappings in the semantically inconsistent affix condition were also randomly intermixed within the training session (see Fig. 3 for a graphical representation of this training regime). The learning phase ended with a free recall test conducted exactly as in Experiment 2.

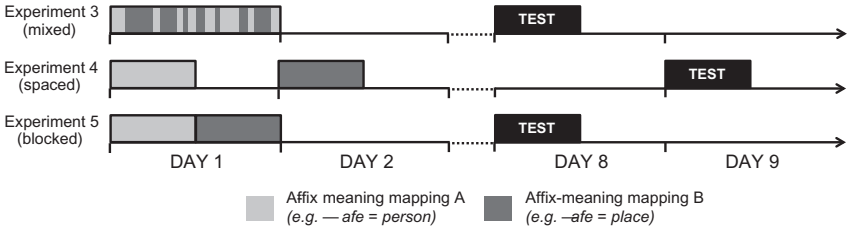

Fig. 3. Diagram illustrating the training regimes of semantically inconsistent affixes used in Experiments 3–5. In Experiment 3 novel words using the two affix-meaning mappings were randomly intermixed, in Experiment 4 they were blocked and separated by 24 h, in Experiment 5 they were blocked with no intervening consolidation opportunity.

4.1.3.2. Testing phase. Participants completed the sentence congruency and recognition memory tasks in exactly the same manner as in previous experiments.

4.2. Results

4.2.1. Training

To investigate whether participants learned words in both semantic consistency conditions equally well, we analysed performance in the last block of the recall task in the training session. Participants recalled an average of 89% (±1.49%) of the words with a semantically consistent affix, and 86% (±1.49%) of the words with an inconsistent affix. A logistic mixed-effects model with no random slopes showed this difference to be significant, $z = 2.30$, $p = .02$.

4.2.2. Free recall

At the end of training, participants recalled on average 41% (±1.71%) of the words with a semantically consistent affix, and 37% (±1.71%) of the words with an inconsistent affix. A logistic mixed effects model (with no random slopes) showed a trend-level difference, $z = 1.78$, $p = .07$, between the conditions.

4.2.3. Sentence congruency

As before, incorrect responses (2.4% of the data) and extremely long RTs (above 1500 ms; 4.4% of the data) were removed. Congruency (congruent vs. incongruent) and semantic consistency (consistent vs. inconsistent) were included as fixed factors. No random slopes were included. The factor of congruency was not significant, $\chi^2(1) = 1.66$, $p = .20$, and neither was semantic consistency, $\chi^2(1) = 0.11$, $p = .74$. However, the interaction between the two factors was significant, $\chi^2(1) = 5.19$, $p = .02$. This interaction reflected a significant congruency effect in the semantically consistent condition, $\chi^2(1) = 6.14$, $p = .01$, but no congruency effect in the semantically inconsistent condition, $\chi^2(1) = 0.46$, $p = .50$. These data are depicted in Fig. 4A.

One participant's data were excluded from the analysis of accuracy scores in the congruency decision component of the sentence congruency task for responding "does not make sense" on every trial. Accuracy in both the semantically consistent (75 ± 4.68%) and inconsistent conditions (59 ± 2.47%) was above chance (both $ps < .001$). Semantic consistency was included as a fixed factor in the analysis, and random slopes for the effect of consistency were included as they significantly improved the model fit. A significant difference was observed between these conditions, $z = -5.68$, $p < .001$.

4.2.4. Recognition memory

Data from the recognition memory task are presented in Fig. 4B. An ANOVA on the *d*-prime values with consistency and knowledge type revealed a significant main effect of consistency, $F(1,23) = 5.03$, $p = .04$, and of knowledge type, $F(2,46) = 167.21$, $p < .001$. The interaction between the two was not significant ($p = .88$).

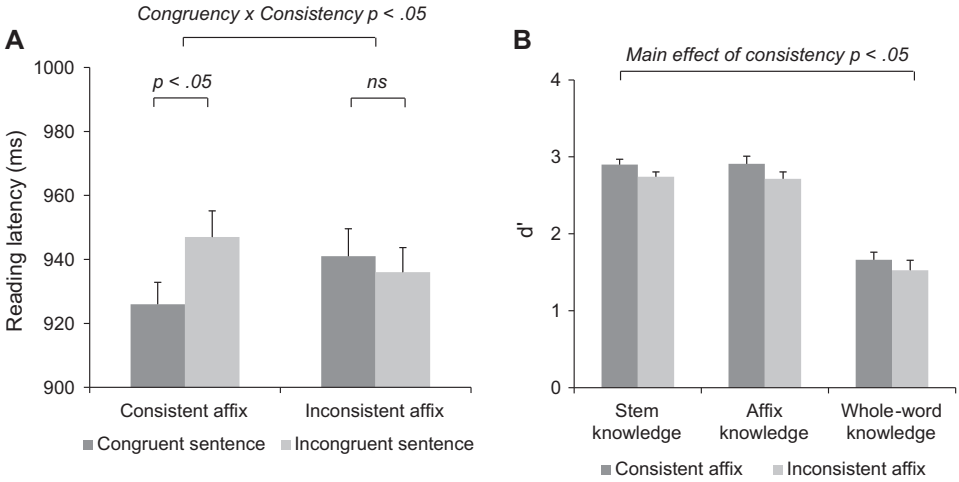

Fig. 4. (A) Reading latencies in Experiment 3 to words in semantically congruent or incongruent sentence contexts for affixes in the semantically consistent and inconsistent conditions. (B) *d*-prime scores for the different knowledge types in the recognition memory task for semantically consistent and inconsistent newly learned affixes. Error bars represent within-participant standard error of the mean. ns = not significant.

4.3. Discussion

Experiment 3 investigated the impact of semantic consistency on affix learning and generalisation. When affixes carried more than one meaning, this reduced participants' ability to make explicit congruency judgments, although their performance remained significantly greater than chance. However, while we observed a significant congruency effect on reading aloud for the semantically consistent affixes (replicating Experiments 1 and 2), this effect was significantly greater than the null effect of congruency observed for semantically-inconsistent affixes, indicating an absence of generalisation for the inconsistent affixes in this speeded task.

Episodic memory appears also to have been affected by semantic consistency. In the free recall task there was a trend-level effect towards better recall for newly learned words with consistent affixes. In the recognition memory task participants were overall more accurate with stimuli in the consistent condition. There was also a numerically small but statistically significant 2% advantage in the last block of the recall task in training for words with affixes in the semantically consistent condition. These data suggest that semantic consistency has a pervasive impact across many domains of language. As outlined in the introduction to this experiment, semantic consistency affects both the processing and the representation of various linguistic units. Here we show that semantic consistency may also, at least under the current training regime, determine whether newly learned linguistic representations (affixes in our case) can generalise, and that consistency affects the strength of episodic representations that begins to emerge already during training.

5. Experiment 4

The failure of semantically inconsistent affixes to generalise in Experiment 3 in the speeded reading aloud task is potentially problematic given that semantic inconsistency is a common feature of language and morphology (e.g., the affix *-er* in *wilder* and *teacher*, as described earlier). The very same problem is encountered in associative learning. In studies of associative learning, participants are asked to learn paired stimuli (A–B; e.g., pairs of words). If after learning a study list (A–B) participants are asked to learn an interference list where the first member of the original pair is now paired with a new stimulus (A–C), memory for the study pairs becomes severely impaired, a form of catastrophic interference (e.g., Bower, Thompson-Schill, & Tulving, 1994).

CLS accounts of memory, however, offer a potential solution. These accounts suggest that new information can be added to established neocortical memory without the kind of catastrophic interference seen in associative learning if the new information is added gradually and interleaved with presentations of the old information. These conditions are met by holding new information back from the neocortex at the time of learning, and allowing the information to be added to the neocortex offline during consolidation in a more gradual manner. The prediction, then, is that catastrophic interference can be avoided by allowing the initial information (A–B) to consolidate prior to adding the new overlapping information (A–C). A number of studies into associative learning have shown this to be the case: interference can be avoided if a consolidation opportunity (best served by a period of sleep) is allowed before training on the interference (A–C) pairs (Drosopoulos, Schulze, Fischer, & Born, 2007; Ellenbogen, Hulbert, Jiang, & Stickgold, 2009; Ellenbogen, Hulbert, Stickgold, Dinges, & Thompson-Schill, 2006; Sheth, Varghese, & Truong, 2012).

Based on these predictions we hypothesised that generalisation of semantically inconsistent newly learned affixes should be possible if we allow the first affix-meaning mapping to consolidate before training participants on the second affix-meaning mapping. Therefore we trained participants on the same items as in Experiment 3, but spread over two days. Participants first learned new affixes, all of which referred consistently to one semantic category. After a 24-h consolidation opportunity, participants returned and now received more training on the same affixes. Importantly, for half of the affixes the semantic category had now been changed (i.e. the affixes were now being made semantically inconsistent) while for the other half the semantic category remained the same as it had been the previous day (i.e. these affixes remained semantically consistent). This way we could ensure that, unlike in Experiment 3, in the semantically inconsistent affixes one affix-meaning mapping had been allowed to consolidate before introducing the second mapping (see Fig. 3).

5.1. Method

5.1.1. Participants

Twenty-four native English-speaking participants were recruited (18 female, 5 left-handed, mean age = 20). All were students at Royal Holloway, University of London, and paid for their participation.

5.1.2. Materials

The materials in all phases and tasks were exactly the same as in Experiment 3.

5.1.3. Procedure

The learning phase was completed over two consecutive days followed by the testing phase one week after the second training session. The same equipment was used as in previous experiments.

5.1.3.1. Learning phase. As in Experiment 3, participants learned a total of 64 novel words: 32 in the semantically consistent condition (eight stems per affix) and 32 in the semantically inconsistent condition (eight stems per affix). However, unlike Experiment 3, training in this experiment involved two sessions conducted over two successive days. On the first day, participants learned 32 novel words using 8 new affixes (4 stems per affix). Half of these words were in the semantically consistent condition, and half were in the semantically inconsistent condition (Table 4). However, the exemplars used in both of these conditions always used a consistent meaning for each affix. On the second day, participants learned the other set of 32 novel words, using the same affixes as in the first session. However, in the semantically inconsistent condition, now half of the affixes were associated with a different meaning category than in the previous session (i.e. they *became* semantically inconsistent as a result of training in this session), while the other half were associated with the same meaning category (i.e. they *remained* semantically consistent). Participants were not told or warned about the introduction of semantic inconsistency.

On both training days, all words were encountered 11 times, once in each of eight typing task blocks, and once in each of three recall blocks (order of blocks was the same is in Experiment 3). Each session lasted roughly 45 min and the presentation order within each block was randomised. Both sessions ended with the free recall task.

5.1.3.2. Testing phase. The test phase was identical to Experiment 3 in all respects.

5.2. Results

5.2.1. Training

To investigate whether participants learned words in both semantic consistency conditions equally well, we analysed performance in the last block of the recall task in both training sessions. On the first day, participants recalled an average of 88% (±3.02%) of the words in the semantically-consistent condition, and 89% (±1.78%) of the words in the semantically-inconsistent condition (although at this point in time these affixes were also still consistent and would become inconsistent only with further training on the following day). On the second day, participants recalled an average of 88% (±2.35%) of the words in the semantically-consistent condition, and 84% (±2.26%) of the words in the semantically-inconsistent condition. A logistic mixed-effects model with the fixed factors of consistency and session but no random slopes showed a marginal main effect of session, $z = -1.94$, $p = .053$, reflecting the small decline is recall on the second day. However, there was no main effect of semantic consistency ($p = .22$) or interaction between semantic consistency and session ($p = .48$).

5.2.2. Free recall

One participant's free recall data were lost due to experimenter error. At the end of the first training session, the remaining participants recalled on average 61% (±3.78%) of the words in the semantically-consistent condition, and 59% (±3.96%) of the words in the semantically-inconsistent condition. At the end of the second training session, they recalled 53% (±3.68%) of the words in the semantically-consistent condition, and 50% (±2.72%) of the words in the semantically-inconsistent condition. A logistic mixed effects model with the fixed factors of consistency and session, and with random slopes for session, showed a significant main effect of session, $z = -2.31$, $p = .02$ reflecting lower recall after the second session. However, there was no main effect of semantic consistency ($p = .28$) or interaction with session ($p = .66$).

5.2.3. Sentence congruency

Incorrect responses (9.8% of the data) and extremely long RTs (above 1500 ms; 5.5% of the data) were removed. Congruency (congruent vs. incongruent) and semantic consistency (consistent vs. inconsistent) were included as fixed factors. No random slopes were included. The factor of congruency was significant, $\chi^2(1) = 10.84$, $p < .001$, but semantic consistency was not, $\chi^2(1) = 1.66$, $p = .20$. There was also no interaction between the factors, $\chi^2(1) = 0.01$, $p = .91$. The significant effect of congruency coupled with the absence of an interaction suggests that the congruency effect was present in both semantic consistency conditions. To confirm this conclusion, we evaluated the effect of congruency separately in the two consistency conditions. A significant effect of congruency was found both in the consistent, $\chi^2(1) = 6.58$, $p = .01$, and inconsistent conditions, $\chi^2(1) = 4.71$, $p = .03$. These data are depicted in Fig. 5A.

One participant's data were excluded from analysis of accuracy scores in congruence judgement for responding "does not make sense" in nearly every trial. Levels of accuracy for the remaining participants were very similar to those in Experiment 3 both in the semantically consistent (76 ± 4.57%) and inconsistent conditions (59 ± 2.93%). Semantic consistency was included as a fixed factor in the analysis, and random slopes for the effect of consistency were included as they significantly improved the model fit. This analysis revealed a significant effect of semantic consistency, $z = -5.76$, $p < .001$. Nonetheless, accuracy scores in both conditions were significantly higher than chance (both $ps < .001$).

5.2.3.1. Congruency effect associated with meanings learned on day 1 vs. day 2. In order to establish whether the congruency effect in reading aloud seen in the semantically inconsistent affixes was observed both for the meaning category acquired on day 1 and for the category acquired on day 2 we restricted the analysis to inconsistent affixes and coded the congruent trials separately as being congruent with respect to the meaning learned on day 1 or on day 2. To see whether this added factor of day interacted with the overall congruency effect, we compared the magnitude of the congruency effect across the two day conditions using the generalised linear hypothesis test implemented in the

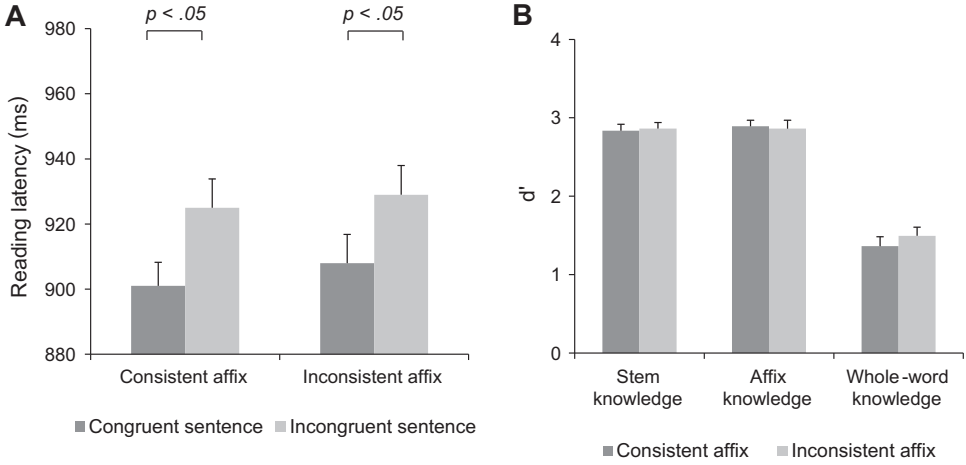

Fig. 5. (A) Reading latencies in Experiment 4 to words in semantically congruent or incongruent sentence contexts for affixes in the semantically consistent and inconsistent conditions. (B) *d*-prime scores for the different knowledge types in the recognition memory task for semantically consistent and inconsistent newly learned affixes. Error bars represent within-participant standard error of the mean.

multcomp package in R. No such difference was found, $z = 1.46$, $p = .16$. As the congruency effect is not modulated by day of learning, these data suggest that both meanings associated with an affix give rise to a congruency effect.

5.2.4. Recognition memory

The recognition memory data are shown in Fig. 5B. An ANOVA on the *d*-prime scores showed no effect of semantic consistency, $F(1,23) = 0.28$, $p = .60$, but did show a main effect of knowledge type, $F(2,46) = 191.36$, $p < .001$. No significant interaction was observed ($p = .21$).

5.2.5. Comparison of the sentence congruency effect between non-spaced learning (Experiment 3) and spaced learning (Experiment 4)

Data from Experiments 3 and 4 were combined to establish whether a statistically reliable difference in the sentence congruency effect could be observed between the two experiments. These comparisons are summarised in Fig. 6. The semantically consistent condition was analysed first, with congruency and experiment included as fixed factors. No random slopes were included. No interaction between the factors was observed, $\chi^2(1) = 0.01$, $p = .94$, while the main effect of congruency did reach significance, $\chi^2(1) = 12.70$, $p < .001$, confirming that the congruency effect in the consistent condition was equivalent across the two experiments. In contrast, in the inconsistent condition we did find a significant interaction, $\chi^2(1) = 4.17$, $p = .04$, while no significant main effects were observed. This interaction confirms that the difference in the congruency effect in the inconsistent condition across the two experiments is statistically reliable.

5.3. Discussion

Experiment 4 explored one potential method for avoiding the disrupting effect of semantic inconsistency on generalisation by allowing the first meaning of a novel affix to consolidate before introducing a second, inconsistent meaning for the same affix. This training manipulation was modelled on methods that have been shown to be effective in avoiding interference between inconsistent pairs in associative learning experiments (Drosopoulos et al., 2007; Ellenbogen et al., 2006, 2009; Sheth et al., 2012). This manipulation was successful: we now found a significant semantic congruency effect

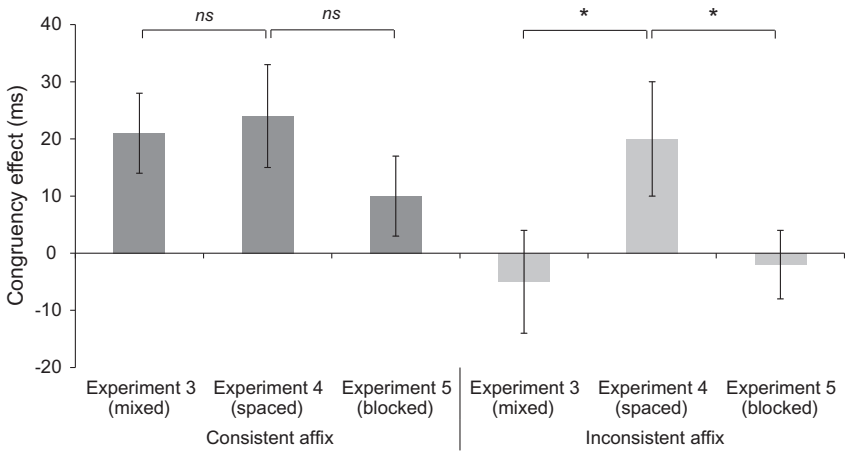

Fig. 6. Magnitude of the sentence congruency effect in Experiments 3–5, in the semantically consistent and inconsistent conditions. Error bars represent between-participant standard error of the mean. $^{*}p < .05$, ns = not significant.

on reading aloud for both the consistent and inconsistent conditions. Furthermore, we were able to establish that in the inconsistent affixes, both of the two meanings afforded semantic congruence effects, demonstrating that these affixes had truly been associated with two different meanings. We also showed that in the inconsistent affixes, the sentence congruency effect was significantly larger in the current experiment than in Experiment 3, while no such difference was observed (or predicted) in the consistent condition.

In the explicit congruence judgement task both conditions were significantly above chance, although participants were again more accurate with consistent affixes. In Experiment 3 we saw poorer decision performance in the same condition where we failed to see a congruency effect in reading latency (i.e. the semantically inconsistent condition). In Experiment 4 we saw exactly the same pattern in judgement performance but this did not stop the emergence of a congruency effect in reading latency. This dissociation reinforces the view that these two tasks reflect very different cognitive processes to achieve generalisation; we will return to these mechanisms in Section 7.

6. Experiment 5

We have argued that semantically inconsistent affixes in Experiment 4 were successfully generalised because consolidation was allowed to operate before introducing the second affix-meaning mapping, while in Experiment 3 generalisation was disrupted because the two mappings were presented simultaneously in the same training session. However, the opportunity for consolidation was not the only difference between the two experiments: it is possible that the critical factor was that the introduction of meanings was *blocked* in the experiment. In Experiment 3 exemplars of both mappings were intermixed; in Experiment 4 one mapping was extensively trained first, followed by extensive (and exclusive) training of the second mapping. While CLS accounts might propose that it is consolidation that makes the difference, an episodic account might conversely propose that stimulus blocking (even without consolidation) would help to prevent interference between the two affix meanings by providing an additional context that allows learners to keep the two meanings separate. To test these predictions, in Experiment 5 we trained participants on the same stimuli as in Experiment 4, with the only difference being that we eliminated the consolidation opportunity by training the first affix-meaning mapping first, immediately followed by training of the second mapping on the same day (see Fig. 3).

6.1. Method

6.1.1. Participants

Thirty-two native English-speaking participants were recruited (24 female, 2 left-handed, mean age = 22). All were students at Royal Holloway, University of London, and paid for their participation.

6.1.2. Materials

The materials in all phases and tasks were exactly the same as in Experiments 3 and 4.

6.1.3. Procedure

The learning phase was completed in one session on day 1, followed by the testing phase one week later. The same equipment was used as in previous experiments.

6.1.3.1. Learning phase. The learning phase was identical to that of Experiment 4 except that the two sessions which were separated by a day in Experiment 4 were now carried out in one session without a break in between. Consistent with Experiment 4, participants were not told or warned about the introduction of semantic inconsistency in the second half of the training session. The session lasted roughly 90 min and the presentation order within each block was randomised. The session ended with the free recall test.

6.1.3.2. Testing phase. The test phase was identical to Experiments 3 and 4 in all respects.

6.2. Results

6.2.1. Training

To investigate whether participants learned words in both semantic consistency conditions equally well, we analysed performance in the last block of the recall task in the two halves of the training session. At the end of the first half participants recalled an average of 92% (±1.49%) of the words in the semantically-consistent condition, and 93% (±1.38%) of the words in the semantically-inconsistent condition (although the inconsistency had not been introduced yet). At the end of the second half participants recalled an average of 90% (±1.44%) of the words in the semantically-consistent condition, and 90% (±1.59%) of the words in the semantically-inconsistent condition. A logistic mixed-effects model with the fixed factors of semantic consistency and part of session (first vs. second half) (no random slopes were included) showed a main effect of part of session, $z = -2.60$, $p = .009$, reflecting the numerically small decline is recall in the second half.

6.2.2. Free recall

The free recall task was carried out at the end of the training session. We included semantic consistency as well as the part of the training session in which the words were learned as fixed factors. Participants recalled on average 43% (±4.03%) of the words learned in the first half in the semantically-consistent condition, and 45% (±3.78%) of the words in the semantically-inconsistent condition (recall however that the inconsistency had not been introduced in the first half). They recalled 41% (±3.36%) of the words learned in the second half with a semantically consistent affix, and 44% (±3.50%) of the words with an inconsistent affix. A logistic mixed effects model with random slopes for training session half showed no significant effects (semantic consistency $p = .30$, part of session $p = .61$, interaction $p = .79$).

6.2.3. Sentence congruency

Incorrect responses (6.8% of the data) and extremely long RTs (above 1500 ms; 7.3% of the data[2]) were removed. Congruency (congruent vs. incongruent) and semantic consistency (consistent vs.

[2] This percentage is higher than in the other experiments due to two participants contributing a disproportionate number of very slow trials. If these two participants were excluded, the proportion of removed trials would be highly similar to the other experiments (4.9%). The exclusion of these participants would not change the results.

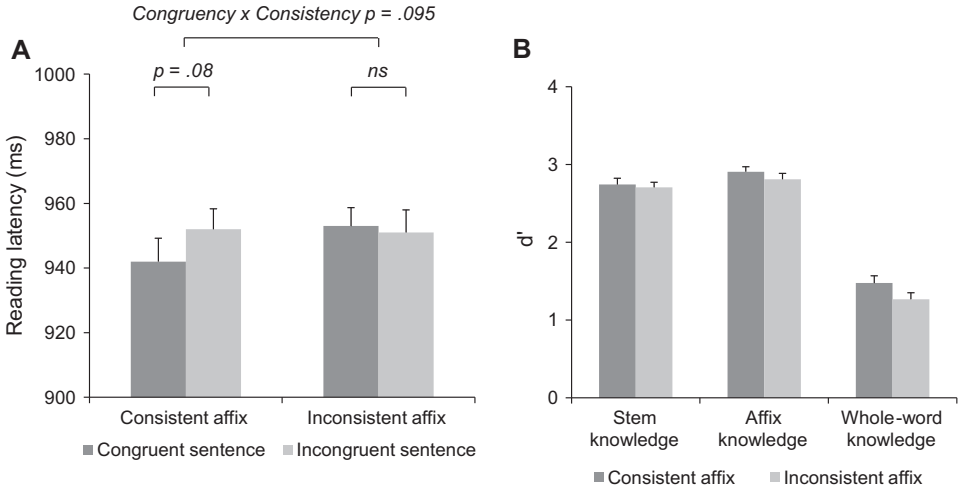

Fig. 7. (A) Reading latencies in Experiment 5 to words in semantically congruent or incongruent sentence contexts for affixes in the semantically consistent and inconsistent conditions. (B) *d*-prime scores for the different knowledge types in the recognition memory task for semantically consistent and inconsistent newly learned affixes. Error bars represent within-participant standard error of the mean. ns = not significant.

inconsistent) were included as fixed factors. No random slopes were included. The factor of congruency was not significant, $\chi^2(1) = 0.71$, $p = .40$, and neither was semantic consistency, $\chi^2(1) = 0.52$, $p = .47$. The interaction between the factors, however, showed a trend-level effect, $\chi^2(1) = 2.78$, $p = .095$. We again evaluated the effect of congruency separately in the two consistency conditions. A marginally significant effect of congruency was found in the consistent condition, $\chi^2(1) = 3.02$, $p = .08$,[3] but not in the inconsistent condition, $\chi^2(1) = 0.29$, $p = .59$. These data are depicted in Fig. 7A.

Accuracy scores in the congruence judgement component of the sentence congruency task were analysed as before. Two participants' data were excluded from this analysis for responding "does not make sense" on every trial. Again, accuracy rates were similar to the previous two experiments in the semantically consistent (77 ± 4.26%) and semantically inconsistent conditions (65 ± 2.74%). Semantic consistency was included as a fixed factor in the analysis, and random slopes for the effect of semantic consistency were included. A significant effect of semantic consistency was observed, $z = -5.30$, $p < .001$, though accuracy scores in both conditions were significantly higher than chance (both *p*s < .001).

6.2.3.1. The sentence congruency effect associated with meanings learned in the first half of the training session vs. second half. To establish whether potential congruency effects in the semantically inconsistent affixes could be observed for the meaning category acquired in the first half or for the category acquired in the second half, the analysis was restricted to inconsistent affixes and the congruency effect in the two meaning categories was compared in the same way as in Experiment 4. The congruency effect did not differ between the two categories, $z = -0.45$, $p = .65$, suggesting that the effect in semantically inconsistent affixes was absent both for meanings learned in the first and for meanings learned in the second half of the training session.

[3] It is difficult to judge with how much confidence one can reject the null hypothesis when marginally significant effects are involved. Therefore we calculated the Bayes factor associated with the congruency effect here, using the Bayes calculator of Dienes (2011). Based on data from Experiments 3 and 4 which include the same condition of semantically consistent affixes as Experiment 5, we represented our theory (that there should be a difference between reading latencies in congruent and incongruent conditions) as a uniform distribution with a lower bound of 0 (no congruency effect) and an upper bound of 23 (the average magnitude in milliseconds of the congruency effect across Experiments 3 and 4). This resulted in a Bayes factor of 2, meaning that the alternative hypothesis (predicted by our theory) is twice as likely as the null hypothesis (that there is no difference between the conditions).

6.2.4. Recognition memory

Recognition memory data are shown in Fig. 7B. An ANOVA on the *d*-prime scores showed a main effect of knowledge type, $F(2,62) = 383.99$, $p < .001$, but no main effect of semantic consistency ($p = .12$) or an interaction between the two factors ($p = .12$).

6.2.5. Comparison of the sentence congruency effect between spaced learning (Experiment 4) and blocked learning (Experiment 5)

The semantically consistent condition was analysed with congruency and experiment included as fixed factors. No random slopes were included. A significant effect of congruency was observed, $\chi^2(1) = 8.58$, $p < .001$. No interaction between the factors was observed, $\chi^2(1) = 0.64$, $p = .42$, confirming that the congruency effect in the consistent condition was equivalent across the training regimes. In contrast, in the inconsistent condition we did find a significant interaction between congruency and experiment, $\chi^2(1) = 3.87$, $p = .049$, while no significant main effects were observed (Fig. 6). The interaction confirms that the difference in the congruency effect in the inconsistent condition across the two experiments was statistically reliable.

6.3. Discussion

The only difference between Experiment 4 and the current experiment was that here we removed the consolidation opportunity that in Experiment 4 followed the acquisition of the first affix-meaning mapping in semantically inconsistent affixes. In Experiment 5 we saw that the consequence of this was that the sentence congruency effect in reading aloud was lost for these affixes. Statistical comparison of the effect across the two experiments confirmed the change was significant (Fig. 6). As in the previous four experiments, we found a congruency effect for semantically consistent affixes (although in this experiment it reached trend-level statistical significance only, it was not statistically different from the congruency effect in Experiment 4). Performance in the explicit congruence judgement task was perfectly consistent with the previous two experiments: both consistency conditions were above chance, but accuracy in the semantically consistent condition was higher.

7. General discussion

Information that can be retrieved from memory ranges from detailed representations of individual events (e.g., it rained yesterday) to general knowledge that has been extracted from an accumulation of multiple individual events (e.g., it is likely to be sunny today given that it is summer and we find ourselves in the south of England). While this is an uncontroversial observation, the nature of the representations that support general knowledge and the processes by which those representations are acquired are far from clear. For example, it is possible that general knowledge is extracted and stored in the form of abstract representations that exist alongside episodic representations (e.g., McClelland et al., 1995). Alternatively, it might be that there is no need for storing abstract representations and that general knowledge can be computed when needed by retrieving and averaging multiple episodic representations (e.g., Hintzman, 1986, 1988) to produce a temporary abstraction. We argue that by mapping the processes involved in the *acquisition* of new general knowledge one can better understand the nature of the representations that underlie this knowledge and the processes involved in generalisation.

We conducted five experiments in which adult participants learned novel affixes embedded in meaningful novel words (e.g., *buildafe*, *sleepafe*, *teachafe*). Following training, we assessed generalisation of the meanings of the novel affixes to previously-untrained exemplars (e.g., *sailafe*), using two measures taken from a sentence reading task. One measure assessed how well participants could judge whether the sentence made sense; this measure forced participants with little time pressure to compute the meanings of the untrained exemplars. The other measure assessed whether there was a sentence congruency effect in speeded reading aloud. This measure did not *require* participants to compute the meanings of the untrained exemplars, but we reasoned that if they had established an affix representation, then a semantically-appropriate sentence frame should benefit reading aloud.

Alongside these measures, we assessed participants' episodic memory for the novel words they had learned, using a standard recognition memory task.

The five experiments reported assessed how these two measures of generalisation were influenced by three variables: (a) the opportunity for memory consolidation between training and testing; (b) the contextual diversity of the novel affixes (i.e. the number of unique stems to which they attach); and (c) the consistency with which the novel affix modifies the meaning of the stem. In all cases, these variables had differential effects on the two generalisation tasks. In respect of the explicit congruency judgment task, performance was significantly above chance in all cases, and was unaffected by (a) the opportunity for consolidation between training and test; and (b) the number of exemplars assigned to each novel affix in the training set. These data from the explicit congruency judgments indicate that participants gained sufficient experience from *all of our training conditions* to compute the meanings of the novel affixes in unfamiliar context (i.e. to generalise their knowledge of the meanings of the novel affixes). In contrast, evidence of a semantic congruency effect on speeded reading aloud *was restricted to particular training conditions*, namely, (a) when there was an opportunity for consolidation between training and test; (b) when novel affixes were paired with a sufficient number of different exemplars during training; and (c) when novel affixes altered the meanings of stems in a semantically-consistent manner, unless (d) the presentation of different meanings was broken by a consolidation interval. Outside of these narrow conditions, no semantic congruency effect on reading aloud was observed, suggesting that in this task untrained exemplars such as *sailafe* were given insufficient time to be processed in a meaningful way (i.e. participants were not generalising their knowledge of the meanings of the novel affixes).

We believe that these findings pose serious challenges for any theory of generalisation based on a single mechanism. As we introduced at the outset, functionally distinct episodic and abstractionist single-mechanism theories have been proposed to explain how we acquire general knowledge from exposure to a limited set of specific instances. These two classes of theory differ most clearly when considering whether and how cognitive processes involved in initial acquisition contribute to later generalisation. Episodic single-mechanism accounts propose that initial learning is achieved by laying down representations of single instances or episodes. These episodes are then combined *during retrieval* to support generalisation (e.g., Hintzman, 1986, 1988). By these accounts, then, generalisation to novel instances involves additional cognitive operations that blend multiple instances of relevant learned items so as to generate the appropriate response. Such a theory provides no explanation however for why factors such as consolidation, contextual diversity, and semantic consistency should operate differently on different indices of generalisation. For example, why does the opportunity for consolidation between training and test modulate generalisation in the reading aloud task but not in the explicit judgment task? Similarly, why is generalisation in the explicit judgment task unaffected by contextual diversity within the training set, and yet low contextual diversity blocks generalisation in the reading aloud task?

Abstractionist accounts propose that abstract, non-veridical representations are generated during initial learning (e.g., Posner & Keele, 1968). In these accounts, generalisation is not achieved during retrieval but by calling on representations of the general structure of particular sets of instances which develop during initial acquisition. Again, it is difficult to understand how a single, abstractionist learning mechanism could lead to different patterns of generalisation in different tasks.

While our findings pose a challenge for purely episodic or purely abstractionist accounts of generalisation, we believe that they are highly compatible with dual-mechanism accounts that combine episodic and abstractionist representations. In the following, we therefore lay out a theory of language learning and generalisation based in particular on the complementary learning systems framework proposed by McClelland et al. (1995).

7.1. Two mechanisms that enable generalisation

One of the unique insights derived from the experiments reported here is that a comprehensive theory of generalisation must take into account the specific cognitive demands of the situation in which generalisation arises. Specifically, there appears to be a fundamental difference in the way that generalisation is achieved in situations comprising online language comprehension and production

tasks (as in our reading aloud task) as compared to situations in which one is required to deliberately and purposefully, with little time pressure, to make a decision about a newly encountered stimulus based on previously encoded knowledge (as in our explicit congruency judgement task). We refer to the former type of generalisation as "online generalisation", to emphasise the notion that this type of generalisation occurs during the course of automatic language processing without deliberate effort. Conversely, we refer to the latter type as "offline generalisation". The diverging results from our two different generalisation tasks suggest that the tasks reflect these different forms of generalisation, and that they may rely on different representations that code the information acquired during training. In the following we consider what we can learn about the nature of these representations from the series of experiments presented here.

7.1.1. Offline generalisation

As outlined above, offline generalisation was largely unaffected by consolidation, family size, and semantic inconsistency. We believe that these observations can be accommodated by a representational architecture that makes use of non-overlapping episodic representations that are available immediately after training.

How might generalisation arise from distinct episodic representations? Kumaran and McClelland (2012) recently developed a computational model of generalisation that makes theoretical predictions about immediate generalisation (based on episodic memory only) that fit elegantly with our data. They noted that there are several demonstrations of immediate generalisation (i.e. generalisation that does not require consolidation to occur), and that generalisation in these circumstances appears to be hippocampally mediated. To simulate the neural mechanisms underlying this immediate generalisation, Kumaran and McClelland's (2012) model of generalisation retains the largely non-overlapping structure of hippocampal episodic representations, but is able to generalise by storing both a veridical memory trace (e.g., a paired associate A–B) linked to its individual components (e.g., A and B) and by allowing bidirectional activation of these two representational layers. However, while these mechanisms are consistent with our notions of offline generalisation, we would argue that these mechanisms are not sufficient to support online generalisation.

Indeed, Kumaran and McClelland (2012) themselves point out that although the hippocampus is able to generalise in this manner, it is likely to support generalisation only over limited timescales and in situations in which all relevant information is hippocampally represented. The hippocampus has only limited storage capacity (Treves & Rolls, 1994), and semantic knowledge is eventually consolidated to the neocortex which stores a more stable and longer-lasting representation of semantic similarity structure. Thus the most effective forms of generalisation for online language processing should be supported by a different class of representations.[4]

7.1.2. Online generalisation

The impact that consolidation, family size, and semantic consistency have on speeded online generalisation of novel affixes suggest that this type of generalisation is supported by representations that emerge gradually over time and rely on an architecture of overlapping representations that are shared among multiple affixed words. Overlapping representations allow the discovery of shared structure amongst a set of items, and critically, the use of this shared information for the purposes of generalisation. The latter conclusion is motivated by our findings from Experiment 2 in which only affixes learned in a sufficient number of overlapping novel words lead to online generalisation. The

[4] A reviewer suggested that if participants became explicitly aware of the meanings of the affixes during training, this explicit knowledge alone might be sufficient to perform the offline generalisation task. Thus, the reviewer suggested that it may be unnecessary to propose operations such as blending of episodic traces (e.g., Hintzman, 1986, 1988) or spreading activation across multiple representational layers (Kumaran & McClelland, 2012). One way to evaluate this possibility would be to ask participants at the end of the training to produce a "translation" for each affix. The current experiments did not include such assessment. However, Tamminen et al. (2012) reported data from a related task, 2AFC definition selection, where participants immediately after training showed highly accurate but very slow explicit judgements about affix meanings. While future studies should evaluate explicit knowledge in more detail during and immediately after training, the effortful nature of getting to this information suggests that explicit knowledge is not readily available immediately following training and requires additional time-consuming cognitive operations.

requirement for overlapping representations can also explain the failure of online generalisation for affixes with inconsistent meanings in Experiment 3. Experiments 4 and 5, however, show that consolidation provides one critical mechanism by which overlapping but inconsistent affix representations can by learned (by allowing one affix-meaning mapping to stabilise during consolidation before introducing a conflicting mapping).

Why is consolidation so critical for establishing overlapping but not distinct representations? The CLS account of memory suggests that this reflects distinct computational specialisations of two different forms of learning: the neocortex is a "slow learner"; hence initial learning is achieved by episodic memory systems (in the hippocampus) that encode more rapidly. The neocortex needs to adopt a slow learning rate due to the nature of the information it stores. Unlike the hippocampal system which specialises in the encoding of individual learning episodes, the neocortex needs to discover the consistent meaning of each of the novel affixes that is learned. This requires knowledge of the form to meaning structure inherent in the entire ensemble of learning episodes (in our case, the ensemble constitutes all the individual trained words, and the relevant structure is the semantic information shared by all words that include a specific affix). According to the CLS theory discovery of the shared structure requires repeated interleaved activation of the entire ensemble of learning events in order to allow all the learning events to be reflected in weights that link form and meaning representations. This form of extended, interleaved exposure can be most efficiently achieved after training, during offline consolidation. This claim is consistent with the data from Experiment 1 in which online generalisation was observed a week after training but not on the same day as training.

Finally, it is worth considering why overlapping representations might allow online generalisation when distinct representations do not seem to do so. We suggest this is largely due to processing time constrains. It may be that the computation of generalised information from episodic representations involves time consuming operations, such as the activation and interaction of multiple representational layers as in the Kumaran and McClelland (2012) model, or the averaging of multiple episodic traces as in the Hintzman (1986, 1988) model. On the other hand, overlapping neocortical representations directly encode the information that is to be generalised, and accessing these representations may be faster than computing generalised representations anew at the time of every retrieval occasion.

7.2. From distinct to overlapping representations

Taken together, the data from the five experiments presented here support a view of language learning and generalisation that makes use of both episodic (distinct) and abstract (overlapping) representational systems. Critically, we have presented evidence of a qualitative change in speed and efficacy of generalisation that occurs before and after memory consolidation. Online generalisation in our speeded reading task emerged only after a period of offline consolidation. In line with CLS accounts, we therefore propose that efficient generalisation is achieved by overlapping representations in neocortical systems. In generating overlapping representations we follow McClelland et al. (1995) and others in proposing a role for consolidation to solve the problem of catastrophic interference (French, 1999; McCloskey & Cohen, 1989) where the adding of new information into overlapping neocortical representations disturbs existing information. Consolidation allows gradual, interleaved learning at the neocortical level, thus avoiding catastrophic interference with existing knowledge as well as enabling the entire ensemble of learning events, rather than just the most recent individual event, to contribute to the discovery of shared structure in the mapping from affix form to affix meaning.

The neural processes that operate during consolidation are becoming increasingly well understood. During consolidation the hippocampus is thought to replay episodic memories encoded earlier (O'Neill, Pleydell-Bouverie, Dupret, & Csicsvari, 2010), allowing temporally controlled extended reactivation of the neocortical memory trace, thus avoiding massed exposure of new knowledge that might otherwise lead to catastrophic interference. Instances of language learning related to generalisation have indeed been shown to involve hippocampal–neocortical interaction during consolidation. Neural markers of hippocampal–neocortical dialogue that occurs during sleep (e.g., sleep spindles and slow wave activity) have been associated with integration of new words in phonological and semantic

neocortical memories (Tamminen, Lambon Ralph, & Lewis, 2013; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010) and evidence from fMRI and MEG shows overnight changes in cortical responses to novel spoken words consistent with the operation of consolidation mechanisms (Davis, Di Betta, Macdonald, & Gaskell, 2009; Gagnepain, Henson, & Davis, 2012; Takashima, Bakker, van Hell, Janzen, & McQueen, 2014). Further, consistent with a theoretical framework in which consolidation supports the establishment of overlapping representations that support more effective generalisation, there is some evidence that sleep-dependent consolidation facilitates abstraction of new grammar in infants (Gómez, Bootzin, & Nadel, 2006) as well as the generalisation of new phonetic learning in adults (Fenn, Nusbaum, & Margoliash, 2003).

Although we have identified consolidation after learning as a critical process for the strengthening of neocortical representations and emerging generalisation, it is worth noting that the CLS account does not suggest that consolidation is the only means to achieve neocortical learning. In the preceding paragraph we described spontaneous reactivations of hippocampal memory traces which likely occur mostly during sleep. However, spontaneous reactivations can be replaced by continuing training trials (which can be thought of as externally rather than hippocampally driven reactivations). For example, amnesic patients with extensive hippocampal damage are able to learn new semantic information, and even generalise this to some degree, but require a larger number of learning trials over an extended period of time to achieve this compared to control subjects (e.g., Hamann & Squire, 1995; Knowlton & Squire, 1993). It is therefore possible (and an interesting avenue for future work) that we could observe generalisation in our affix learning paradigm in the absence of a consolidation opportunity if we significantly increased the number of training trials and spaced these over a day (see Lindsay & Gaskell, 2013, for a similar demonstration in a spoken word learning paradigm which however did not test generalisation).

7.3. Relationship between our artificial stimuli and natural morphological word formation processes

Before considering how our results might impact on models of lexical processing, it is important to consider whether morpheme learning in our experiments is sufficiently similar to natural morphological acquisition to support conclusions about lexical processing. One key question is whether the morphemic regularities in our stimuli are like those that arise in natural language. Morphological word formation comprises both derivational affixation and compounding. The key difference between these processes is that derivational affixation involves attaching a bound form (i.e. one that cannot stand alone) onto an existing root, while compounding involves combining independent word forms into a single form. Derivational affixation can further involve attaching an affix to a bound root (e.g., *–mit*, as in *submit, permit*) or to free-standing stems (e.g., *dark* in *darkness,* or *kind* in *kindness*).

These different types of derivation vary in the extent to which the meanings of the whole form can be predicted from the meanings of the parts (i.e. compositionality). In English, derivational affixation as applied to existing stems (e.g., *darkness, kindness*) tends to be highly compositional (e.g., *–ness* forms abstract nouns that almost always refer to the quality or state of the adjective used as the stem), while derivational affixation as applied to bound stems (e.g., *submit, permit*) or compounding tends to carry more idiosyncratic information. For example, while the compounds *snowman*, *milkman*, *chairman*, and *fireman* all relate in some way to a person, the relationship between the whole form and the constituents are different in each case (e.g., a snowman is a man-like figure made of snow, but a milkman is a person who delivers the milk each morning). However, while derivational affixation as applied to existing stems tends to be more compositional than the other two varieties of morphological word formation, it also allows idiosyncrasies (e.g., consider the affix *–ist*; a cyclist is someone who engages in the act of cycling; a harpist is someone who plays the harp; a geologist is someone who specialises in geology; a Calvinist is someone who adheres to the doctrine of Calvin; a racist is someone who believes that one race is superior to another; a stylist is someone who arranges objects, clothes, or food in a stylish way).

In order to facilitate learning and to discourage explicit hypothesis testing during learning about the meanings of the novel affixes, we created semantically-rich definitions for our novel words which tended to convey more idiosyncratic information than is the case for the most compositional types of derivational affixation. Indeed, some of our novel affixes seem somewhat like compounds in the way

in which they convey meaning (e.g., –*afe* in *sleepafe*, *teachafe*, *buildafe* resembles compounds using 'man'). However, unlike in the case of compounds, participants in our experiments are never exposed to the right-hand constituent (i.e. the affix) in isolation, which is the critical feature of compounding. Thus, we suggest that our stimuli are more like derivational affixes than compounds. However, we would also argue that which type of morphological word formation process more closely resembles our artificial stimuli does not really matter for the conclusions that we draw. That is, none of our conclusions would be affected if future research were to show that simpler or less idiosyncratic affixes could be more readily learned.

One final point concerns the fact that all of our experiments used suffixation, which raises the question of whether the effects observed would extend to prefixes. While this question can only be settled by further experiments directly comparing the acquisition of novel suffixes vs. prefixes, we predict that acquiring new prefix representations might be more difficult and require more extensive training than acquiring new suffix representations. This is mainly motivated by the suffixing preference, the tendency to prefer suffix morphology over prefix morphology across the languages of the world (e.g., Cutler, Hawkins, & Gilligan, 1985). Artificial language learning studies have shown that this preference may be due to suffixes being more informative cues to language structure than prefixes (St. Clair, Monaghan, & Ramscar, 2009), further suggesting that language users might more readily acquire new suffixes than prefixes.

7.4. Implications for models of lexical processing

Having established that our artificial stimuli closely resemble existing morphological formations, we now consider the implications of our results for the further development of lexical processing models. The majority of localist accounts of word recognition propose that lexical representations are morphemically structured (e.g., Marslen-Wilson et al., 1994; Taft, 1994), or that morphemically-structured representations are accessed in parallel with whole word representations (Caramazza et al., 1988; Schreuder & Baayen, 1997). Both types of architecture include abstract localist representations of morphemic units that never occur in isolation (such as affixes). Yet, none of these models makes any specific claims concerning how these abstract morphemic units might be acquired. The experimental data reported in this paper therefore provide important constraints on how these models could be developed further to model the process that gives rise to abstract affix representations sufficient to support generalisation. We have shown that the morphemic abstraction assumed by localist models of lexical processing occurs only after consolidation and not during initial encoding. Further, the development of these abstract representations would appear to depend on the acquisition of a sufficient number of semantically-consistent affixed forms. One critical challenge for these models will therefore be to offer some functional explanation as to why these factors appear to underpin this abstraction process.

Distributed-connectionist models of lexical processing, on the other hand, appear to make more specific proposals about the mechanisms supporting acquisition of morphemically-structured representations (e.g., Plaut & Gonnerman, 2000; Rueckl & Raveh, 1999). In these accounts, morphemic representations are encoded in hidden units that mediate the mapping between the forms of words and their meanings. These representations are acquired by the operation of an error-correcting neural-network learning algorithm (typically back-propagation, though other algorithms would likely lead to similar outcomes; see Plaut & Shallice, 1993, for simulations). These learning algorithms generate affix representations by virtue of the systematic relationships between the orthographic or phonological form of specific affixes (-*er*), and the meanings of affixed forms (e.g., *dancer, teacher* and *thinker* all refer to people). Unlike localist accounts, however, these representations are graded as a function of the compositionality and frequency of the relationship between affixed forms and their meanings, and so will emerge more strongly when affixes arise in multiple words in the training set and modify the meanings of stems in a highly-consistent manner. One interesting characteristic of these models, then, is that although they are trained on whole words, they come to acquire structured, componential representations that act as abstract representations of morphemic units in supporting generalisation. Thus, neural network learning algorithms *appear* able to acquire abstract morphemic representations without requiring anything more than learning a sufficient number of

morphologically related words, with sufficiently transparent form-to-meaning correspondences (see Plaut & Gonnerman, 2000, for illustrative simulations of how changes to the structure of the network's vocabulary can impact on internal representations of morphemic units).

However, although existing distributed models of morphological processing use learning algorithms to derive morphemic representations, these learning processes are seldom intended as a key part of the theoretical account embodied by the model (Plaut & Gonnerman, 2000; Rueckl & Raveh, 1999). Instead, these networks are tested once learning is complete, and simulations are proposed to capture aspects of the adult system rather than developmental profiles (though, see Plunkett & Marchman, 1993, for a model of inflection acquisition). One challenge for these models in trying to simulate *how* morphological knowledge is acquired is that they can show only a limited amount of new learning (e.g., acquiring a new word or affix) without requiring additional training on pre-existing knowledge (i.e. representations of previously learned words are rendered unstable by training on new words). For example, a model taught a new pseudo-affixed word with high-frequency (such as the word *twitter* referring to a microblogging internet service) might struggle to retain the knowledge that words ending *-er* typically refer to people (as in *dancer*, *teacher*, etc.) unless these existing words were relearned in parallel with acquisition of the new word *twitter*. This is the problem of catastrophic interference, here described in the context of morphemic learning rather than associative learning as initially demonstrated (cf. French, 1999; McCloskey & Cohen, 1989).

In response to this problem, proponents of connectionist learning models have typically invoked the notion of complementary learning systems – specifically, a division between rapid medial temporal/hippocampal and slower neocortical learning – in order to achieve a more appropriate trade-off between stability of existing knowledge and plasticity in acquiring new knowledge (McClelland et al., 1995). Several aspects of our data demonstrate that learning in human participants may be subject to similar limitations as learning in distributed-connectionist models. Like distributed-connectionist models, human learning and generalisation appears to be highly vulnerable to new information that contains inconsistencies, and effective generalisation appears to require an extended timescale through a process of consolidation. Thus, far from being a reason to argue against distributed connectionist models and to favour localist accounts that lack this trade-off between stability and plasticity (e.g., Page, 2000), it would seem that similar constraints on the speed of learning and efficacy of generalisation are observed in adults learning novel affixes.

There are two ways in which the present results extend our understanding of the contribution of complementary learning systems to morphological processing and language learning. The first is that we see evidence for constrained forms of generalisation immediately after learning. As discussed in connection with the Kumaran and McClelland (2012) model, representations encoded in the hippocampus support some limited forms of generalisation that may be particularly evident in non-speeded tasks which require participants to apply their knowledge to new problems. The second and more interesting implication is that the representations initially encoded into short-term storage in the medial temporal lobe ultimately constrain the degree of generalisation that can be observed after consolidation. This is most apparent in comparing the results of Experiments 3, 4 and 5. These studies showed that inconsistent affix meanings learned on the same day fail to generalise even after consolidation, but that they do generalise if learning of the two meanings arises on separate days. Thus, inconsistencies interfere with processes that discover shared structure in the training set through overnight consolidation. In this way, episodic storage of recently acquired novel affixes acts as a bottleneck for learning; only if a sufficient set of consistent affixed meanings are learned and stored in episodic memory do we subsequently observe generalisation of affix representations. In this respect, then, the results of Experiment 2 provide another constraint on this episodic bottleneck – a sufficient number of related forms must be acquired if consolidation is to lead to abstract morphemic representations.

7.5. Distinguishing speeded and non-speeded tasks in language learning

If speeded and non-speeded tasks indeed tap into different forms of generalisation of newly learned affixes, it should be possible to observe this distinction in other language learning paradigms. In fact, a similar distinction emerges as a common theme in a number of studies looking at various

forms of adult word learning. For example, Tamminen et al. (2010) trained participants on spoken novel words and found that in a speeded 2AFC recognition memory task participants' reaction times were significantly facilitated by a sleep-dependent period of overnight consolidation. Dumay and Gaskell (2007) on the other hand found no overnight change in accuracy rates in a non-speeded version of the same task (although their participants were close to ceiling). It is possible that the lack of consolidation effects in Dumay and Gaskell (2007) was because participants could make the non-speeded 2AFC decision on the basis of episodic memory, while speeded responses made by participants in Tamminen et al. (2010) were facilitated after consolidation by neocortically stored lexical representations.

A similar conclusion can be reached by considering word learning data reported by Tamminen and Gaskell (2013). Here participants learned novel written words and their definitions (e.g., that *feckton* is a type of cat). Tests of memory for the meanings included non-speeded cued recall of the novel word meanings ("what does *feckton* mean?") and speeded lexical decision to familiar words primed (e.g., *kitten*) or unprimed (e.g., *bat*) by the recently learned novel words (e.g., *feckton*). While the non-speeded recall task revealed excellent memory of the meanings immediately after training (which declined over time after training), the novel word primes began to influence lexical decision reaction times to familiar targets only after a period of consolidation. It is likely that participants were able to retrieve the novel word meanings from episodic memory immediately after training, given a task with no time constraints. However, semantic priming effects may require lexical knowledge that overlaps for new and existing words in order to support priming of familiar words like *kitten*. Hence, the presence of such overlap was only revealed when measured by speeded, online language processing tasks (speeded lexical decision in this case). Data such as these from a range of word learning paradigms support our conclusion that speeded tasks tap into gradually emerging representations of newly learned language, while non-speeded tasks can draw upon episodic information encoded immediately during learning.

7.6. Conclusions and implications for pedagogy

We have presented evidence of two different forms of generalisation that are served by different types of representations. We suggest that one form of generalisation is possible based on episodic memory immediately after learning, possibly through a mechanism of reactivation and blending of hippocampally-stored representations like the one proposed by Kumaran and McClelland (2012), and analogous to that proposed in episodic accounts. However, this mechanism for generalisation is severely constrained in several ways. For one, it is constrained by the time pressure on the forms of generalisation that are required for comprehension and production of natural language. There is likely to be insufficient time for the extensive computations required for hippocampal generalisation during online language processing. Furthermore, since the hippocampus appears to have only a limited capacity for retaining a longer-term store of language knowledge there may be a further limitation of the range of circumstances, topics, or words over which this offline generalisation may be apparent. For example, it might be that hippocampal generalisation is particularly vulnerable when learners fail to retain sufficiently strong memories of key items. Therefore a second mechanism is needed which allows instant access to generalised information and has less restrictive capacity limitations. We suggest that this mechanism takes the form of overlapping neocortical representations that emerge gradually during consolidation. Together these two forms of generalisation ensure that newly acquired learning episodes can be used immediately to guide behaviour and decision making but the integrity of existing semantic memory is protected by being learned at a slower rate during memory consolidation.

We further conclude that the operation of these mechanisms should be carefully considered in teaching language and literacy. Learners in these domains are expected to do more than encode discrete pieces of information. Indeed, the expectation is that learners will be able to extract something more general from the whole of these discrete pieces of information. For example, by teaching beginning readers the pronunciations of words such as *clown, down, gown, frown*, teachers may have the expectation that pupils will learn something more general about the pronunciation of the trigraph -*own*, even if this principle is not taught explicitly. Our findings together with our notion of the episodic

bottleneck suggest clear strategies for enhancing generalisation in these kinds of situations. Specifically, one would want to ensure that there were a sufficient number of exemplars presented for learning and that these were highly consistent, so that inconsistent items like *blown* would be introduced only following consolidation of the general principle being taught.

Similar constraints apply in learning to write morphological elements which is widely recognised as important for accurate spelling (Bowers, Kirby, & Deacon, 2010; Nunes & Bryant, 2006). For example, if children were being taught how to spell the *-ing* inflectional morpheme through the presentation of multiple exemplars (e.g., *jumping*, *kicking*, *playing*, *showing*), it would be important to ensure that teaching materials do not include cases in which the same *-ing* ending appears in a non-morphemic context (e.g., *ceiling, pudding, morning*). Furthermore, it might be that common orthographic alterations of the stems (e.g., *swimming*, *dancing*) should also be excluded during initial learning of simple stem + *ing* items. Our findings suggest that including such cases in a single lesson or set of teaching materials may have the effect of blocking generalisation of the *–ing* affix that is being taught. Rather than including small numbers of exceptional items in a single lesson, it would be better to group items which include highly-systematic orthographic alterations into lessons taught on subsequent days which focus exclusively on specific orthographic changes such as e-deletion (in *dancing, racing*) or consonant duplication (in *swimming, running*). By grouping these items in this way it would prevent these items from being learned as exceptions and instead allow learners to acquire and generalise all of the relevant principles involved in the spelling of the *-ing* affix.

Many theoretical and applied questions remain unanswered. For example, is generalisation in the classroom enhanced by making the principle being taught explicit in the form of simple, verbalisable rules (as in the use of phonics instruction for learning to read)? If so, what is the mechanism that underpins that advantage? Similarly, what is the role of sleep (if any) in the acquisition of general knowledge? If there is a role for sleep in the acquisition of general knowledge, then what are the pedagogical implications of poor sleep behaviour? If there are any very general conclusions to be drawn from our work, one is certainly that there are great opportunities for engagement between scientific theory and educational practice in language learning and generalisation. The other very strong conclusion is that theories of language processing will be substantially enhanced by taking seriously the problem of acquisition, in such a way that engages rigorously with the state-of-the-art in memory research.

Acknowledgments

This research was funded by an Economic and Social Research Council (ESRC) grant to KR and MHD (RES-062-23-2268). JT was also supported by a British Academy Postdoctoral Fellowship and MHD by the UK Medical Research Council (MC_US_A060_0038). We are grateful to Professor Alec Marantz for advice regarding the selection of our stimuli. We also thank Liz Shepherd and Amy Gatton for assistance in running the experiments and scoring the data.

Appendix A

See Tables A.1–A.5.

Table A.1
Accuracy rates in the recognition memory task in Experiment 1 (between-participant standard error in parenthesis).

	Item type	Time of test	
		Immediate	Delayed
% correct	Trained stem, trained affix	93% (±1.24%)	93% (±1.15%)
	Untrained stem, trained affix	99% (±0.36%)	87% (±2.33%)
	Trained stem, untrained affix	99% (±0.61%)	98% (±1.28%)
	Recombinant words	79% (±2.96%)	64% (±3.65%)

Table A.2
Accuracy rates in the recognition memory task in Experiment 2 (within-participant standard error in parenthesis).

Item type	Family size	
	Large family	Small family
Trained stem, trained affix	84% (±3.40%)	95% (±3.50%)
Recombinant words	65% (±3.60%)	67% (±4.86%)

Table A.3
Accuracy rates in the recognition memory task in Experiment 3 (within-participant standard error in parenthesis).

	Item type	Semantic consistency	
		Consistent	Inconsistent
% correct	Trained stem, trained affix	85% (±3.50%)	83% (±2.66%)
	Untrained stem, trained affix	97% (±1.65%)	97% (±1.41%)
	Trained stem, untrained affix	98% (±1.45%)	97% (±1.63%)
	Recombinant words	68% (±3.90%)	67% (±3.96%)

Table A.4
Accuracy rates in the recognition memory task in Experiment 4 (within-participant standard error in parenthesis).

	Item type	Semantic consistency	
		Consistent	Inconsistent
% correct	Trained stem, trained affix	86% (±3.24%)	86% (±3.72%)
	Untrained stem, trained affix	96% (±1.31%)	96% (±1.24%)
	Trained stem, untrained affix	97% (±1.55%)	96% (±1.54%)
	Recombinant words	56% (±4.59%)	60% (±4.83%)

Table A.5
Accuracy rates in the recognition memory task in Experiment 5 (within-participant standard error in parenthesis).

	Item type	Semantic consistency	
		Consistent	Inconsistent
% correct	Trained stem, trained affix	86% (±2.31%)	84% (±2.31%)
	Untrained stem, trained affix	95% (±1.09%)	96% (±1.31%)
	Trained stem, untrained affix	98% (±1.03%)	98% (±1.05%)
	Recombinant words	61% (±3.54%)	58% (±3.10%)

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*, 814–823.
Algeo, J. (1991). *Fifty years among the new words: A dictionary of neologisms*. Cambridge, UK: Cambridge University Press.
Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, 91*, 7041–7045.
Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.
Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.
Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homophony, and productivity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*, 489–511.
Bower, G. H., Thompson-Schill, S., & Tulving, E. (1994). Reducing retroactive interference: An interference analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 51–66.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. *Review of Educational Research, 80*, 144–179.
Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review, 121*, 248–261.
Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York: Academic Press.
Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition, 28*, 297–332.
Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing*. London: Academic Press.
Cutler, A., Hawkins, J. A., & Gilligan, G. (1985). The suffixing preference: A processing explanation. *Linguistics, 23*, 723–758.
Davis, M. H., Di Betta, A., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21*, 803–820.
Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society, 364*, 3773–3800.
De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes, 15*, 329–365.
Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.
Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General, 136*, 169–183.
Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*, 35–39.
Ellenbogen, J. M., Hulbert, J. C., Jiang, Y., & Stickgold, R. (2009). The sleeping brain's influence on verbal memory: Boosting resistance to interference. *PLoS ONE, 4*, e4117.
Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: Sleep, declarative memory and associative interference. *Current Biology, 16*, 1290–1294.
Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*, 614–616.
Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language, 63*, 117–130.
Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35*, 116–124.
French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3*, 128–135.
Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in human auditory cortex. *Current Biology, 22*, 615–621.
Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*, 105–132.
Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431–436.
Gómez, R. L., Bootzin, R., & Nadel, L. (2006). Naps promote abstraction in language learning infants. *Psychological Science, 17*, 670–674.
Hamann, S. B., & Squire, L. R. (1995). On the acquisition of new declarative knowledge in amnesia. *Behavioral Neuroscience, 109*, 1027–1044.
Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411–428.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.
Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262*, 1747–1749.
Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review, 119*, 573–616.
Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 608–622.
Love, B. C., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.
Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology, 23*, 2023–2027.
Marcus, G. F., Pinker, S., Ullman, M. T., Hollander, M., Rosen, T. J., & Xu, F. (1992). *Overregularization in language acquisition. Monographs of the society for research in child development 57 serial no. 228*. Chicago, IL: University of Chicago Press.
Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences, 262*, 23–81.
Marslen-Wilson, W., Ford, M., Older, L., & Zhou, X. (1996). The combinatorial lexicon: Affixes as processing structures. In G. W. Cottrell (Ed.), *Proceedings of the eighteenth annual conference of the cognitive science society*. NJ: Lawrence Erlbaum.
Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*, 3–33.
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*, 419–457.
McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation, 24*, 109–165.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*, 295–322.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
Merkx, M., Rastle, K., & Davis, M. H. (2011). The acquisition of morphological knowledge investigated through artificial language learning. *The Quarterly Journal of Experimental Psychology, 64*, 1200–1220.
Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 275–292.
Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review, 110*, 611–646.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–61.
Nunes, T., & Bryant, P. (2006). *Improving literacy through teaching morphemes*. London: Routledge.
O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology, 10*, 58–70.
O'Neill, J., Pleydell-Bouverie, B., Dupret, D., & Csicsvari, J. (2010). Play it again: Reactivation of waking experience and memory. *Trends in Neurosciences, 33*, 220–229.
O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences, 6*, 505–510.
Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*, 443–467.
Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes, 15*, 445–485.
Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10*, 377–500.
Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21–69.
Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.
Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods, 39*, 859–862.
Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1083–1095.
Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes, 23*, 942–971.
Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285–308.
Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., et al (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition, 40*, 1095–1108.
Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*, 245–266.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
Rueckl, J. G., & Dror, I. E. (1994). The effect of orthographic-semantic systematicity on the acquisition of new words. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance, XV*. Hillsdale, NJ: Erlbaum.
Rueckl, J. G., & Raveh, M. (1999). The influence of morphological regularities on the dynamics of a connectionist network. *Brain and Language, 68*, 110–117.
Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. II). Cambridge, MA: MIT Press.
Schreuder, R., & Baayen, R. H. (1997). How simplex complex words can be. *Journal of Memory and Language, 37*, 118–139.
Sheth, B. R., Varghese, R., & Truong, T. (2012). Sleep shelters verbal memory from different kinds of interference. *Sleep, 35*, 985–996.
St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*, 1317–1329.
Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9*, 271–294.
Takashima, A., Bakker, I., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Richness of information about novel words influences how episodic and semantic memory networks interact during lexicalization. *NeuroImage, 84*, 265–278.
Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition, 125*, 107–112.
Tamminen, J., & Gaskell, M. G. (2013). Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology, 66*, 1001–1025.
Tamminen, J., Lambon Ralph, M. A., & Lewis, P. A. (2013). The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *Journal of Neuroscience, 33*, 15376–15381.
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*, 14356–14360.
Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209–253.
Treiman, R., & Cassar, M. (1996). Effects of morphology on children's spelling of final consonant clusters. *Journal of Experimental Child Psychology, 63*, 141–170.
Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus, 4*, 374–391.
Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology, 66*, 319–327.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press.
Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71–86.
West, R. F., & Stanovich, K. E. (1978). Automatic contextual facilitation in readers of three pages. *Child Development, 49*, 717–727.
Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society, 17*, 766–780.
Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 1160–1173.
Zwitserlood, P., Bölte, J., & Dohmes, P. (2000). Morphological effects on speech production: Evidence from picture naming. *Language and Cognitive Processes, 15*, 563–591.