Sleep–related consolidation of new form–meaning mappings: the acquisition of arbitrary and systematic mappings in adult language learning

Anastasia Roshchupkina

Doctor of Philosophy

University of York
Psychology

September 2016

Abstract

Six experiments investigated the role of sleep-related memory consolidation in learning new words (e.g. *tib bisesh* = queen). We tested the predictions of the Complementary Learning Systems model (CLS; McClelland et al., 1995) that sleep-related consolidation varies with arbitrariness in the form-meaning mapping. New determiners (*tib, ked*) and suffixes (*-esh, -ool*) systematically mapped on to the referents' natural gender. Stem-meaning mappings (e.g. *bis-* = queen, *jor-* = cowboy) were arbitrary, as the meaning of the stem could not be predicted from its phonology. In Experiments 1 and 2 there was one determiner and two suffixes per gender (*tib* = female, *ked* = male; *-esh, -eem* = female, -*ool, -aff* = male). In Experiment 2 overnight polysomnography data was collected, to correlate slow wave sleep (SWS) with arbitrary mapping recall. In Experiments 3 and 4, there were two determiners and one suffix per gender (*tib, paz* = female, *ked, jov* = male; -*eem* = female, *-ool* = male). In Experiments 5 and 6, the systematic mapping included the suffixes only (without determiners), and the number of exemplars was increased in Experiment 6. The memory for the arbitrary mappings was tested in recall and recognition tasks. The knowledge of the systematic mappings was tested in generalisation tasks. As an exploratory investigation of Ullman's Declarative/ Procedural model (e.g. Ullman, 2001) we also correlated measures of arbitrary and systematic mappings with a declarative and a procedural task. As predicted by the CLS, there was evidence to suggest that sleep was beneficial for the memory of the arbitrary mappings, but not for the systematic mappings. Determiners required full systematicity to be extracted. Suffixes required increased exemplar variability and no determiners present to be extracted. The findings will be discussed in the context of models of memory consolidation in word learning.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would primarily like to thank my supervisors Dr. Jelena Mirkovic and Professor Gareth Gaskell for their patience, guidance and comments, for which I am infinitely grateful. Also, a separate thank you to Dr. Jelena Mirkovic for her invaluable feedback which has throughout the years inspired me to think about my findings more deeply.

I also thank my TAP panel for useful meetings, especially Professor Andy Ellis who initially encouraged me to embark on this PhD, and was also my supervisor until he retired. I thank all of the SLAM lab members for insightful discussions. For help with polysomnography training I thank Justyna, Jelena and Scott. In addition, I thank Woody Chau for helpful research assistance. Finally, I thank Matthew Rahtz and Dominic Arnold for recording the spoken versions of the language stimuli used in the Experiments.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.
The work within this thesis was completed solely by the author under the supervision of Dr Jelena Mirkovic and Professor Gareth Gaskell.

Data from Experiments 1 and 2 was presented at:
Roshchupkina, A., Mirkovic, J., & Gaskell, G. Sleep-Related Consolidation of New Form-Meaning Regularities. Poster presented at Studying language in mind conference, Birmingham, June 2014.

Chapter 1: The role of memory consolidation in learning new individual memories and overlapping regularities

1.1 Introduction

The aim of this thesis is to investigate the learning of new form-meaning mappings in adults, i.e. the way the sounds of words map on to their meanings. More specifically, we are focusing on the similarities and differences in learning and consolidation of arbitrary vs. systematic aspects of the mappings. Existing natural human languages have mappings which are both arbitrary (e.g. word meanings) and systematic (e.g. grammatical gender implemented via determiners and suffixes), such as those found in the French translations for male singer, '*le chanteur*', and female singer, '*la chanteuse*' (whereby '*chant_*' is the arbitrary stem, '*le*' and '*la*' are the male and female gender determiners respectively, and '*_eur*' and '*_euse*' are the male and female gender suffixes respectively).

Within the framework of this thesis, language will be used as a tool to explore the learning of mappings with varying degrees of systematicity. In addition, the process of consolidation (processes related to learning, remembering and retrieving newly learnt mappings following a delay which contains sleep or the equivalent time awake) will be examined. For this purpose, literature regarding the role of sleep in language learning will be reviewed to examine how sleep affects the consolidation of novel words (i.e. specific individual items), and to provide an introduction to the emerging literature on sleep-associated consolidation. Then, a new dimension – that of the distinction between arbitrariness and systematicity (as found in language mappings) will be introduced. Issues relevant to this dimension will be reviewed, which in the existing memory literature have been addressed in the form of schema formation (i.e. the creation of new mental representations). More specifically, schema formation will be addressed in the form of rule abstraction, generalisation (ability to apply newly learnt abstracted rules to new items, e.g. Gomez, Bootzin & Nadel, 2006) and relational memory (creation of a single coherent schema based on relations between separately learnt individual items, e.g. Lau, Alger & Fishbein, 2011) and discussed in the framework of consolidation effects.

Moreover, an exploratory analysis will be included in this thesis, whereby it will be explored whether the arbitrary and systematic language mappings can be linked to aspects of Ullman's declarative/procedural model (e.g. Ullman, 2001). For this purpose, literature will be presented to suggest that arbitrary language mappings might be linked to the declarative memory system (e.g. Breitenstein et al.,2005), whereas systematic language mappings might be linked to the procedural memory system (e.g. Brill-Schuetz & Morgan-Short, 2014).

In addition, as we are exploring the links between the language mappings and the aspects of the declarative/procedural model (e.g. Ullman, 2001), all of our Experiments will include not only the language tasks, but also a declarative and procedural task. The language, declarative and procedural tasks will be all learned in one session, and all retrieved in one separate, second session. To address this, literature will be presented to suggest that firstly as a result of interference, the effects of sleep-dependent consolidation may be altered (e.g. Barsky, Tucker & Stickgold, 2015). More specifically, the typically seen post-sleep gains on procedural tasks (e.g. Walker et al., 2002), may no longer be observed as a consequence of interference (Barsky, Tucker & Stickgold, 2015). Meanwhile, sleep may protect declarative memory from interference (e.g. Deliens, Leproult, Neu & Peigneux, 2013).

Moreover, literature will be reviewed to suggest that interference may arise during both acquisition (e.g. Brown & Robertson, 2007), and retrieval (Albouy et al., 2008). The consequence of this interference is that the way the information is processed may be different from when information is acquired and retrieved without interference. For example, across memory systems, the learning of a second task may disrupt the consolidation of the task that was learned first (Brown & Robertson, 2007).

1.2 Role of memory consolidation in language learning

The emerging literature suggests that there is an active involvement of sleep in language learning. Gaskell and Dumay (2003) first showed that newly learned words require time to be integrated into the mental lexicon with existing words. This line of studies tested integration by teaching participants new phonological forms, i.e. new spoken words (such as '*cathedruke'* which are based on existing real words such as '*cathedral*', thus sharing a phonological uniqueness point '*cathedr_*'). The new words were close phonological neighbours of existing words and once integrated into the mental lexicon (dictionary, containing information on each word's semantic and phonological properties, e.g. Bates & MacWhinney, 1987) with existing words should slow the recognition of the existing 'base' words (deciding if the word is real or not in a lexical decision task) due to the shared phonological uniqueness point with newly acquired words, a process termed lexical competition.

In Gaskell and Dumay's (2003) set of three experiments participants were trained using phoneme monitoring (on each trial participants saw one of six target phonemes e.g. */d/* on screen and decided if the phoneme was present in each new spoken word). The tests consisted of free recall (participants had 3 minutes to say aloud as many of the trained words as possible), and recognition (each trained word e.g. *'cathedruke'* was presented with a similar untrained word e.g. *'cathedruce'* and participants decided which was familiar). Across the set of experiments, integration was tested using either a lexical decision task in the first two experiments (e.g. hearing *'cathedral'* with *'cathedruke'* and deciding which word is real) or a pause detection task in the third experiment (requiring participants to say whether they heard a pause in each word e.g. *'cathedr_al',* lexical competition in this task would be evidenced by participants taking longer to detect a pause in an existing word such as *'cathedr_al'* if the phonologically similar word *'cathedruke'* has been integrated). Across the set of experiments, participants showed good recognition of newly learnt spoken words immediately following learning and after a delay of both twenty-four hours and seven days. Evidence for the newly learnt words having been integrated into the existing mental lexicon, measured by lexical competition, was seen following three days and seven days after initial learning but not immediately, suggesting that the integration process requires time to develop.

The first direct evidence that sleep (above just the passage of time) plays a role in language learning and consolidation comes from a study by Dumay and Gaskell (2007). This study investigated lexilisation, i.e. the integration of new spoken words into the mental lexicon of pre-existing words. This study showed evidence not only for the active role of sleep in language learning but also for the role of sleep in integration in adults. Participants learnt new spoken words (e.g. *'shadowks'* which was based on the existing word *'shadow'*) by completing a phoneme monitoring task (as with Gaskell & Dumay, 2003). The learning phase occurred either in the morning (such that the following consolidation delay of 12 hours occurred during daytime wakefulness) or in the evening (such that the following consolidation delay of 12 hours occurred during overnight sleep). Participants were tested using free recall, recognition and pause detection tasks (also as with Gaskell & Dumay, 2003). The tests were administered both immediately after learning and after a consolidation delay (of 12 and 24 hours).

Free recall of the newly learned spoken words was found to improve only following a period of sleep (not the equivalent time awake), suggesting that sleep aids the consolidation of newly learned individual words in their exact form, a process termed stabilization. Moreover, newly learned words showed evidence of integration (as measured by the test of lexicalisation, in this case pause detection, only after a period of overnight sleep. This was true both if the period of overnight sleep occurred immediately i.e. with a test 12 hours following learning or after a day spent awake, i.e. with a test 24 hours following learning. There was no evidence of integration immediately after learning or following the equivalent time awake. This suggests that a period of sleep is necessary for newly learned spoken words to be integrated into the mental lexicon. This is a process of enhancement as the newly learned words are not just consolidated individually and separately from other language knowledge but on the basis of their connections (in this case in phonology) with other, already existing language knowledge. Lexical competition due to the integration of the novel words with the existing words (measured using the same tests as Dumay & Gaskell, 2007) was also observed up to eight months following initial new word acquisition, confirming that the words have been stored into the long term mental lexicon (Tamminen & Gaskell, 2008).

A further study (Tamminen, Payne, Stickgold, Wamsley & Gaskell, 2010) also showed direct links between properties of sleep and novel word learning (also using the same items e.g. *'cathedruke'* and the same tests of free recall, recognition (in this case named old-new categorisation) and lexical decision as previous studies e.g. Gaskell & Dumay, 2003 as well as a cued recall test in which the cue was the initial 2 -3 phonemes of the trained new words). In the Tamminen et al. 2010 study, participants were also trained either in the morning or in the evening and tested immediately and after both a shorter consolidation delay (of either a day awake or overnight sleep) and a longer consolidation delay (of seven days).

Both recognition and recall improved after a period of overnight sleep but not the equivalent time awake (these improvements were noticed in the wake group only at the seven-day retest, with performance on the cued-recall test in the wake group even declining significantly during the shorter consolidation delay). Integration of the newly learned words into the mental lexicon was not observed immediately, and required a consolidation delay to emerge (observed in both the wake and the sleep group participants). Moreover, for participants in the sleep group, increased spindle count (as measured by polysomnography) also correlated with the degree of the lexical competition, suggesting that sleep spindles facilitate the integration of new words into the mental lexicon. Improvements in recognition (faster responses) were more associated with the length of slow wave sleep (SWS). These findings confirm that SWS (a stage of sleep) facilitates the consolidation of declarative memories (e.g. Diekelmann & Born, 2010).

The evidence that sleep is found to play an active role in language learning with adults are also found in children for both stabilization (Brown, Weighall, Henderson & Gaskell, 2012) and integration (Henderson, Weighall, Brown & Gaskell, 2012) of newly learned spoken words. In a study testing stabilization (Brown, Weighall, Henderson & Gaskell, 2012), children (both 7 and 12-years old) showed improved performance on a cued recall task only following a delay (of 24 hours) which included a period of sleep, but not after a brief delay (of 3 or 4 hours) spent awake. Recognition performance (which was good initially compared to poor recall immediately after learning) also improved following the brief wake delay as well as the longer (24 hour) delay including sleep, suggesting that sleep is more beneficial for recall than recognition memory.

In a study testing integration (Henderson et al., 2012), children did not show evidence of integration immediately following learning of new spoken words nor after a 12-hour delay awake. Evidence of integration was only found after a 12-hour period of sleep, suggesting that sleep is necessary for integration of new words into the mental lexicon in children as well.

Similarly, Williams and Horst (2014) tested how consolidation affects preschool children's new word learning using a "shared storybook reading task". Children either heard the same story (read by the experimenter in person) repeated three times or heard three different stories once each (the stories were presented within a book). Two new words (referring to objects, e.g. *'sprock'* – handmixer) were embedded within each story (and illustrated within the book). Children were tested on their knowledge of the new words using a comprehension test at four time-points: immediately after learning, following a short consolidation delay of 2.5 hours (of either a nap if the children usually napped or the equivalent time awake), and after two longer consolidation delays of 24 hours (which included overnight sleep) and seven days. As a control task (testing if the children who heard three different stories were paying attention), children answered plot comprehension questions using two forced-choice answers (both answers were taken from the story which the question referred to). During the word comprehension test, on each trial children saw a page with four pictures (some of the 'wrong answer' pictures were from the stories but not all) and had to point to the target (e.g. if the experimenter said *'sprock'* the child should have pointed to the handmixer picture).

The control task testing plot comprehension showed that children who heard the different stories understood all of them (above chance performance), which showed that any differences following consolidation could not be due to differences in understanding. Children who heard the same story three times performed better on the immediate comprehension test (i.e. learned more words initially) than if they heard three different stories. Both children who heard the same story repeated three times and those who heard three different stories showed similar new word comprehension ability following consolidation but only if they napped during the shorter consolidation delay. For the children who did not nap, performance was still lower even after the 24 hours including overnight sleep. Overall children who napped learned more words than the children who did not nap. These findings suggest that in preschool children an immediate period of sleep following learning aids the consolidation of newly learned words. Even if the new words were encountered fewer times and thus initially learned less well, an immediate nap (but not the equivalent time awake) compensated for this acquisition disadvantage.

There are however studies (e.g. Lindsay & Gaskell, 2013; Szmalec, Page & Duyck, 2012) which show some constraints on the role of sleep in language learning. In an adaptation of Gaskell and Dumay's (2003) study (using the same new spoken words, e.g. *'cathedruke'* and similar tasks) Lindsay and Gaskell (2013) used a different training paradigm (in two of the three experiments) to teach participants new spoken words. The training paradigm involved spaced (distributed) learning: participants were repeatedly exposed to the new words across five different time points (rather than learning the same information all at one time point). The first four sessions occurred every 2.5 hours. The fifth session occurred 24 hours following the last session of the first training day (such that this interval included a period of sleep). It was found that sleep led to improvements in both cued recall and recognition; therefore sleep was beneficial for stabilization.

Integration was measured using a lexical decision task (as with Gaskell & Dumay, 2003). If the new spoken words were learned using spaced learning *and* if this was accompanied by spaced exposure to the existing base words (e.g. repeated exposure to both *'cathedruke'* and its existing phonological neighbour *'cathedral'*), sleep was not required for integration (as evidenced by lexical competition effects being observed on the same day as training without sleep). It is possible that repeated exposure to *both* the new and the existing words strengthened their phonological connections, making wake-associated consolidation sufficient. Therefore, whether or not sleep is required for consolidation of new words may be dependent on the training method (possibly the extent to which the training method results in the strengthening of connections within existing knowledge networks).

The idea that sleep effects may be task specific was confirmed in Szmalec, Page and Duyck's (2012) study, who taught participants new words using Hebb learning. Participants repeatedly saw sequences of new word syllables (called Hebb sequences) presented on screen (e.g. *bi-ki-na* which was derived from the existing Dutch word '*bikini').* During the recall test participants saw nine syllables arranged in a circle and were required to (by clicking on each sequence with the mouse) recreate the order in which they were presented (if participants were unsure of a specific position they clicked a '*question mark'* to indicate an omission). Lexicalisation of the trained words (learned as syllables) was tested after 24 hours using both a pause detection task and a lexical detection task (as with Gaskell & Dumay, 2003 but using Dutch words). At the 24-hour retest, words had been lexicalised (i.e. integrated into the mental lexicon) as evidenced by lexical competition effects (on both tasks testing lexicalisation).

In a second experiment, to test whether sleep affected integration, participants (having learned the sequences either in the morning, i.e. the wake group or in the evening, i.e. the sleep group) were tested immediately, after 12 hours and after 24 hours using the pause detection task only (as with Dumay & Gaskell, 2007). Lexical competition was not observed immediately after training, but was present at the 12-hour retest, remaining at the 24 hour retest. Following the consolidation delay, participants who stayed awake and those who slept performed similarly. These findings suggest that while new words are not integrated into the mental lexicon immediately and require time for integration, in some cases (perhaps depending on how the new words were acquired) sleep is not necessary for this process and just the passage of time is sufficient.

To summarise, these findings show that under certain circumstances sleep plays an active role in language learning in both children and in adults. A period of offline consolidation is necessary for firstly the stabilization of new words (as evidenced by improvements in recall and recognition tasks). Secondly, it is necessary for the enhancement of new words, whereby they are integrated into the existing mental lexicon (as evidenced by competition effects of new words with already existing words). Both the stabilization and enhancement of new word representations in the long term memory emerge following a period of sleep but do not always emerge immediately, or after the equivalent time awake. This suggests that sleep has an active role in these processes, further supported by correlations found on recognition tasks with slow wave sleep (SWS) and on lexical competition with sleep spindle count (Tamminen et al., 2010). However, there are also findings (e.g. Lindsay & Gaskell, 2013; Szmalec, Page & Duyck, 2012) which show that in some cases (which may be task dependent i.e. depend on the training method) wake-associated consolidation may be sufficient and no additional benefits of sleep may be observed.

1.2.1 Schema Formation in Language: Systematicity and Arbitrariness

The new dimension which the present thesis aims to add is the role of schema and extraction of regularities in learning and consolidation of sound-meaning mappings. The way different aspects of language (arbitrary versus systematic) are learned is explored within the connectionist framework. The idea that systematic aspects of language rely relatively more on the neocortical systems and the learning of the arbitrary information on the hippocampal systems originates from the Complementary Learning Systems model, introduced by McClelland, McNaughton and O'Reilly (1995). According to this model, systematic aspects of mappings are acquired through repeated exposure to different exemplars and the abstracted regularities can then be generalised (potentially at the cost of forgetting the individual specific items). Arbitrary mappings are learnt in their exact forms and given that they are arbitrary, i.e. with no overlapping regularities, the mappings cannot be applied to new exemplars.

In this thesis we are interested in discussing not only the learning and consolidation of arbitrary and systematic mappings regardless of sleep, but also in discussing the role of sleep in the consolidation of arbitrary and systematic mappings. Consequently, it is relevant to introduce a sleep-related consolidation model, namely a two-stage model (e.g. Walker & Stickgold, 2010). Walker and Stickgold's (2010) model of "sleep-dependent memory evolution" suggests that the role of sleep in consolidation is two-fold. Sleep plays a role in the stabilization (i.e. strengthening) of information exemplars in their exact forms. Sleep *also* plays a role in the re-organization of these exemplars in order to facilitate generalization to previously unseen exemplars.

In order to facilitate generalization, sleep can be said to "evolve" recently acquired memories. This sleep-dependent process of memory "evolution" can be broken down into three separate processes. During a process termed "unitization", sleep creates a "unitized" schema, on the basis of a shared element from a set of distinct exemplars. During a process termed "assimilation", newly acquired individual exemplars are "assimilated", i.e. integrated with already existing knowledge (e.g. semantic) networks. This process of "assimilation" is applicable to the integration of recently learned phonological word-forms into the already existing mental lexicon (as evidenced by Dumay & Gaskell, 2007). During a process termed "abstraction", overarching hidden regularities can be "abstracted" from a set of exemplars, with the hidden regularities also potentially being brought to the learner's awareness. The stabilization of memories in their exact forms occurs during the first stage of sleep-dependent consolidation. This process may be mediated by slow wave sleep (SWS) and may require just one night of sleep (or even a nap containing SWS). Then, the exact memories may "evolve" via processes relating to the extraction, abstraction and generalization of memories during the second stage of sleep-dependent consolidation. These second stage processes may require several nights of sleep to be carried out. A potential consequence of these second stage processes is the loss of the exact individual forms.

While Walker and Stickgold's (2010) model refers to *general* sleep-dependent memory consolidation, it can also be applied to arbitrary and systematic language mappings. In the context of this model, arbitrary mappings would be consolidated in their exact individual forms (during a single night of sleep, possibly mediated by SWS), whereas systematic mappings (requiring the processes of extraction and generalization of regularities) may require several nights of sleep to be consolidated. A potential consequence of the consolidation of the systematic mappings may be the loss of the individual, arbitrary mappings.

To address the distinction between arbitrary and systematic mappings in the context of the role of sleep in language learning, literature regarding rule abstraction, generalisation and relational memory will be discussed. As within this thesis the questions are tested behaviourally, the focus of the literature reviewed will also be on behavioural studies rather than the neural basis of the processes.

1.2.1.1 Schema Formation: Rule Abstraction

Above just consolidating memories exact to their original form, in cases where newly encountered information contains regularities, such regularities can be abstracted to create new "schemas". This process, which is referred to as "abstraction" (as mentioned above) within integrative models of memory consolidation (e.g. Walker & Stickgold, 2010) is useful so that when new information with similar rules is encountered the new "schemas" (having been integrated with pre-existing "schemas" in the knowledge networks) can then be applied to the new exemplars. Lewis and Durrant (2011) refer to this process (by which information is abstracted during sleep) as "information overlap to abstract" ("iOtA").

According to Lewis and Durrant's (2011) "iOtA" model, newly learned memories are replayed (i.e. reactivated within the neural networks in which they were encoded) during the subsequent sleep period (during slow wave sleep in particular). When related memories with overlapping elements are replayed during sleep, the overlapping elements are strengthened while the individual elements are eventually eroded during global downscaling (of neural synaptic connections). Therefore, the overlapping elements are replayed more often, leading to a stronger memory trace and a new schema being created based on the overlapping elements. Thus, replay of overlapping memories during sleep facilitates the abstraction of overlapping elements (or rules) following sleep.

This idea that sleep facilitates rule abstraction has been examined in several studies. For example, one such study used a sequence learning task to test learning in adults (Nieuwenhuis, Folia, Forkstam, Jensen & Petersson, 2013). The letter sequences (5-12 consonants in length, e.g. '*VXSSVS'*) followed a complex rule schema (and in that sense were 'grammatical': e.g. if the first consonant in the sequence was *'V'* the second consonant was always *'X'* and the third consonant was *'S'* and so on as in the example *'VXSSVS'*). During training, participants' short term memory for the sequences was measured by asking participants to type each sequence they just learned. Participants' extraction of the rules was tested using a classification test: participants decided which sequences were 'grammatical' (e.g. if participants abstracted the rules, they should have classified '*VXSSVS'* as grammatical, and '*VXRSSS'* as not grammatical, as a sequence starting with *'VX'* could not be followed by *'R'*). Participants were tested immediately (15 minutes after learning), or after a delay of 12 hours (of either overnight sleep or daytime wakefulness) or after 24 hours (including sleep). It was found that only participants who slept during the consolidation delay (as evidenced by both the 12 hour and 24 hour retest groups) were able to correct judge a newly learnt sequence as ‘grammatical’ or not. Thus, results suggest that sleep improves the extraction of complex rules from a set of related items in adults.

Similarly, Durrant, Taylor, Cairney & Lewis (2011) used a sequence learning task (a statistical learning paradigm) to show the benefits of sleep for statistical rule (probability) abstraction. Participants were exposed to tone sequences (each tone had one of five different frequencies). The way the tones were combined to make sequences was based on probabilities. For each two previous tones that were heard, a given tone would be played on 90% of the trials. On the remaining 10% of the trials it would be followed by one of four other tones. For example, the two tones ‘*3*’ ‘*3*’ would be followed by ‘*1*’on 90% of the trials, and by other tones (not by ‘1’) on the remaining 10% of the trials. Participants were tested immediately after exposure and also after a consolidation delay which involved either a period of sleep or an equivalent period of daytime wakefulness. The consolidation period of sleep or wakefulness lasted for 12 hours in the first experiment and for 90 minutes (in which case the sleep group was the nap group) in the second experiment. Polysomnography was used during the second nap experiment in order to correlate specific sleep stages with participants’ performance. Participants completed a two-alternative forced choice (2-AFC) test, in which every trial involved hearing two sequences of tones. Neither of the two sequences were heard during exposure, but one of the sequences was consistent with the probabilities during exposure and the other was put together randomly and was therefore not based on the probabilities during exposure. Participants responded (by pressing a button) which sequence was more familiar.

Participants’ performance on the 2-AFC test improved (between immediate and delayed test) significantly more if they slept during the consolidation delay (both for a period of 12 hours overnight and for 90 minutes during the day) than if they stayed awake (for 12 hours or 90 minutes respectively). Participants who had a longer consolidation delay (lasting 12 hours) improved significantly more than those who had a shorter consolidation delay. This effect was due to the sleep and wake group effects combined. Moreover, the findings from the nap experiment showed that these sleep-related improvements were facilitated by slow wave sleep (SWS) as evidenced by positive correlations between the amount of SWS and improvement in performance. These findings confirm that sleep (and SWS in particular) both in the form of a shorter nap and longer overnight sleep (compared to the equivalent time awake) benefits the abstraction of statistical probabilities.

Using the same statistical learning paradigm (but with seven tones instead of five), Durrant, Cairney and Lewis (2012) tested participants either 30 minutes after learning or following 24 hours (which included sleep). Participants listened to the sequences during the exposure phase. During the immediate recall test, participants heard sequences (half of which were the same as during exposure and half were not) and indicated if each sequence was familiar or not. The delayed recall test also included new sequences (either with the same statistical probabilities as during exposure or randomly generated) and again indicated if each sequence was familiar. Rest trials (i.e. without sequences, included such that fMRI could be used to measure baseline activity) were also included.
As with Durrant et al., (2011), the findings showed that sleep (a delay of 24 hours compared to 30 minutes) benefitted participants' ability to discriminate between sequences consistent with the trained probabilities and random sequences. This suggested that sleep benefitted statistical probability abstraction and was again (as with Durrant et al., 2011) facilitated by SWS. Moreover, the findings (from fMRI) showed that before sleep (at the 30 minute retest) neural activity was greater in the medial temporal lobe, while following sleep (at the 24 hour retest) neural activity was greater in the striatum (and auditory regions involved in the processing of tone-sequences). This supported the idea that sleep benefits rule abstraction through neural memory reorganisation (from the hippocampus to striatal areas) to enable more automatic processing of the learned material.

Alongside showing that sleep benefits rule abstraction, previous studies have also shown the benefit of sleep in generating explicit awareness of the embedded rules. The first evidence for this idea showed that sleep promotes insight, by using an altered number reduction task which is thought to rely on the procedural memory system (Wagner, Gais, Haider, Verleger & Born, 2004). In this task, participants used two rules to derive new sequences from old number sequences. These sequences could be derived quicker by using a hidden rule – as participants received no instruction to the existence of this rule (and therefore were not consciously searching for it), the rule was acquired implicitly.

Significantly more participants (twice as many) in the overnight sleep group (who slept for 8 hours) as in the daytime wake group were able to solve the sequences by accessing the hidden rule (as evidenced by abrupt rather than gradual improvement in sequence solving). This benefit of sleep was only observed if initial exposure to the sequences (with the hidden rule) occurred prior to the period of sleep (i.e. if there was initial trained information to be consolidated with sleep). This showed that sleep acts on information which participants have already been exposed to (rather than easing immediate insight of information encountered for the first time *after* sleep occurred). In cases where new information containing hidden rules is learnt prior to a period of sleep, the benefit of sleep is in restructuring the newly learnt information such that the rules can be extracted. There is some further evidence suggesting that this sleep dependent insight gain is related to slow wave sleep spindle activity (e.g. Yordanova, Kolev, Wagner, Born & Verleger, 2012).

These results are supported by another study suggesting that sleep (over the equivalent time awake) facilitates the extraction of implicit rules governing repeated information sequences. This study used a serial reaction time task (SRTT) and a "generation task" (Fischer, Drosopoulos, Tsen & Born, 2006). Two groups of participants were tested: an overnight sleep group, and a daytime wakefulness group. The implicit SRTT required participants to as quickly and accurately as possible type various keys signifying the different cue locations. There was a hidden rule based on which the cue locations were repeated.

Based on the decreasing reaction times to repeated (compared to random) sequences, some (implicit) understanding of rules was gained during training. Results of the explicit generation task (which required participants to predict where a target would appear in a sequence) also showed that before sleeping, participants were unable to explicitly replicate the sequence when requested to, but were able to do so after overnight sleep. Participants who acquired the rules without awareness (on the serial reaction time task) became aware of these rules (as measured by the generation task) only if the delay between exposure and test included a period of overnight sleep, but not the equivalent time of daytime wakefulness. As participants' speed (i.e. reaction times) on the implicit SRTT did not differ between initial exposure and after the consolidation delay, this suggests that the benefit of sleep was not in improving this sequence solving skill but in bringing implicitly learnt information to awareness. This is confirmed by the findings from the generation task during exposure. As participants from both groups (those who would sleep overnight and those who would remain awake during the day during the consolidation delay) were at chance on the generation task immediately following exposure (before the consolidation delay), the findings suggest that the sleep state specifically (rather than just a time delay) allows for memories to be reorganised in such a way that (at least grammatical or sequential) rules initially gained without awareness are brought to awareness.

In another sequence learning study (Wilhelm, Rose, Imhof, Rasch, Buchel & Born, 2012), children aged 8 – 11 years old, and adults aged 18 – 35 years old learnt a motor sequence by typing it repeatedly. Participants had to type the eight-item sequence as quickly as possible, by pressing buttons on a button box which lit up. An additional group of children learned a sixteen-item sequence. This is because the children reached ceiling effects, hence the sixteen-item sequence allowed for improvements in performance to occur. These improvements could be correlated with slow wave activity (SWA) during sleep. Participants were asked to recreate the sequence (using the button-box) during a free recall test. This test measured explicit sequence awareness after a consolidation delay of 10-12 hours. The delay included overnight sleep for participants who were trained in the evening, or the equivalent time of daytime wakefulness for participants who were trained in the morning.

With repeated exposure (during training), all participants' reaction times declined significantly. This suggests that repetition improved implicit knowledge of the sequence. Participants who slept (and children more so than adults), had better performance on the free recall task than those who stayed awake during the consolidation delay. This suggests that sleep facilitates explicit awareness of the structure of the sequences which were learned implicitly. Moreover, this improved performance on the free recall task correlated with increased SWA in both children and in adults. Also, in this task children had significantly more SWA than adults, potentially explaining children's better performance following sleep than adults'. Participants (both children and adults) whose brain activity at retrieval was measured (using fMRI) completed an additional cued recall task. In this task the cue was one of the items in the sequence, and participants had to press the two following buttons. Children showed higher left posterior hippocampus activity during this task than adults. These findings suggest that the reorganisation of memories during sleep which allows for implicitly learned information to be brought to explicit awareness is facilitated by SWA and greater hippocampal activity. Children may be better than adults (following sleep) at gaining explicit awareness due to their increased SWA (during sleep) and hippocampal activity (during retrieval).

To summarise, the above studies have illustrated two ideas. Firstly, when new information such as a sequence is learnt which contains a complex hidden rule (i.e. the rule is acquired 'implicitly' through exposure – that is, without awareness of the rule), a period of offline consolidation involving sleep (compared to the equivalent time awake) has been shown to facilitate schema abstraction, as evidenced by the ability to distinguish between 'grammatical' (i.e. rule-following) and 'ungrammatical' (i.e. random or not rule-following) items or sequences (e.g. Durrant et al., 2011; 2012; Niewuwenhuis et al., 2013). These benefits of sleep to schema abstraction have been shown to be facilitated by SWS and neural reorganisation (from the hippocampus to striatal areas following sleep, Durrant et al., 2012). Secondly, the above studies have demonstrated that sleep may facilitate insight, in the sense that following a period of sleep explicit awareness of a rule which was initially learnt without awareness emerges (e.g. Fischer et al., 2006; Wagner et al., 2004; Wilhelm et al., 2012) and is facilitated by increased SWA and hippocampal activity, particularly in children (more so than in adults) (Wilhelm et al., 2012). This suggests that beyond just benefitting the consolidation of individual items in their exact form, sleep also facilitates the restructuring of new memories such that rules previously acquired without awareness are extracted and even become available explicitly following sleep. When information containing hidden rules is acquired implicitly, sleep plays a role in reorganising the memories of this information such that the hidden rule is extracted and becomes consciously available.

1.2.1.2 Schema Formation: Generalisation

The process whereby the abstracted rules can be applied to new items has been referred to as "generalisation" (e.g. Eichenbaum, 2004) or "multi-item generalisation" (Stickgold & Walker, 2013), as the purpose of the creation of such a single schema is for it to then be generalised, i.e. applied to new exemplars which are consistent with the learned regularities. In order for a newly learnt rule to be applied to new items (the process of generalisation), the said rule must first be extracted from the associations presented during training (e.g. Walker & Stickgold, 2010). Therefore, generalisation is taken as evidence of extraction of regularities.

Evidence for the role of sleep in memory generalization comes from a study using synthetic speech (Fenn, Nusbaum & Margoliash, 2003). Over the course of two training sessions, participants were exposed to a speech stream generated by a speech synthesizer (comprised of consonant-vocal-consonant *(CVC)* words) together with the written word forms on screen. Twelve hours following training, participants were tested on their ability to identify words from the speech stream. More specifically, participants listened to the synthetic speech stream and typed the CVC words which they were able to identify from the stream. Participants were also given a pre-test, prior to training, to assess their ability to identify words. The 12 hour consolidation period included either a period of daytime wakefulness if exposure occurred in the morning, or overnight sleep if exposure occurred in the evening. An additional control group was tested immediately (i.e. without a consolidation delay). The exact set of individual words used in the pretest was different from the sets of words used in the two training sessions and in the retest. This suggests that increased ability of identifying CVC words from the synthetic speech stream would reflect generalization ability (i.e. generalization of 'acoustic-phonetic context' mappings). The immediate test (of the control group) showed that participants' ability to identify the words in the speech improved significantly with training (compared to the pretest before training). Performance improved significantly more if the consolidation delay included sleep than after the equivalent time awake. Findings suggest that the consolidation of generalizable mappings is facilitated by sleep.

Differences between wake and sleep-associated consolidation were also tested in a further study using synthetic speech with generalizable mappings (Fenn, Margoliash & Nusbaum, 2013). Participants learnt new words either through generalisation-training or rote-training. Generalisation-training involved learning 300 words, each repeated once. Rote-training involved learning 20 words (which were used as part of the generalisation-training and on all of the tests) repeated 15 times. During training, both groups heard the words and saw the written word forms presented on screen (all of the words were created using a speech synthesizer). Recognition performance (i.e. ability to identify words) of the participants in both the generalisation-trained and the rote-trained groups declined during the wake delay (as evidenced by a test 12 hours following initial training). Recognition performance was restored only in participants in the generalisation-trained group (and not the rote-trained group) following a subsequent sleep period (as evidenced by a test 24 hours following initial training). This suggests that sleep preferentially consolidates information which was learnt in such a way that it can be generalised (i.e. through exposure to a large set of items, over exposure to a small set of items which is more likely to facilitate consolidation of individual specific items). This is in line with previous studies which suggest that whether sleep benefits consolidation depends on the training method (e.g. Lindsay & Gaskell, 2013; Szmalec et al., 2012).

Also looking at generalisation ability both immediately following learning and after a consolidation delay, Tamminen, Davis, Merkx and Rastle (2012) investigated the generalization of newly learnt affixes. Participants learnt new affixes (e.g. *'_nule'*) added to existing known stems (i.e. word roots such as *'build'*) and the meanings of the new stem-affix combinations (e.g. *'buildnule'* is someone who builds furniture quickly). Participants learned the new words (such as *'buildnule'*) by hearing and seeing them (together with their definitions) on screen. After each trial participants typed the new word. At three points during exposure participants were tested with a recall task: participants saw a definition and typed the new word which it referred to. Immediately or two days following training, participants completed three tests. During the tests, participants were presented with both trained and untrained (appearing for the first time at test) words. The untrained words in all of the tests included '*untrained stem-trained affix*' combinations, for example "sailnule". A speeded shadowing task involved hearing and pronouncing items (as quickly as possible). In this task, there were two additional untrained combinations: '*trained stem- untrained affix*' (e.g. "sail*nule*", as participants were trained on "build*nule*") and '*untrained stem-untrained affix*' (e.g. a previously unseen stem paired with the untrained affix "*nept*").

Performance on the speeded shadowing task (pronouncing new items) showed no benefit for the newly learned affixes immediately after learning but performance improved significantly after a period of consolidation. Therefore, this study suggested that a period of consolidation is required for context-independent generalisation to emerge. This may be because episodic memories are formed on the basis of acquiring specific individual items, while context-independent generalisation may rely on separate abstract representations (Davis & Gaskell, 2009), and may require a longer time period. However, it is important to note that in this study it was not possible to attribute consolidation benefits specifically to sleep, due to the delayed test being run two days after learning.

The idea that a period of sleep improves the abstraction of grammatical rules has also been examined in infants. For example, Gomez, Bootzin and Nadel (2006) exposed

15 month old infants to an auditory stream of an artificial language. Infants' preferred orientation (head turning) and reaction times were measured. The phonological stream consisted of three-word sequences e.g. '*vot-X-rud'*. There was no regularity to predict what '*X'*, i.e. the middle word would be. The first word (e.g. '*vot*') always predicted the last word (e.g. '*rud'*) in the training sequences. At test, sequences either contained the same items as during training or new, untrained items.

Infants who stayed awaked showed preference for trained over untrained items, suggesting that wake-associated consolidation (in infants) is more beneficial for the consolidation of specific individual items. Infants who napped during the consolidation-delay, at the post-nap test showed preference for nonadjacent dependencies (i.e. regularities between the first and third element in the sequence) which were consistent with the nonadjacent dependencies of the first test sequence (regardless of whether the nonadjacent dependencies in the first test sequence were trained or untrained). However, this only occurred if the nap-group infants were exposed to a larger variety of middle elements in the exemplars during training (the experimental nap group was exposed to 24 middle '*X'* elements, while the nap-control group (which showed no evidence of having extracted regularities) was exposed to only 3 middle *'X'* elements during training). The nap-group infants did not show evidence of remembering individual learnt items, unlike the infants who stayed awake. This suggests that only the infants who napped (and not those who remained awake) processed (whether with or without awareness) the information that the sequences they heard contained within them regularities, suggesting that sleep is beneficial for the extraction of regularities and more so than for consolidation of specific individual items.

A second study used the same materials, training and testing procedures (Hupbach, Gomez, Bootzin and Nadel, 2009). Children aged 15 months were exposed to the same auditory input of three-item long phrases as in the previous study (Gomez et al. 2006). Training was followed by either a nap (immediately or in a follow-up experiment within 4 hours of learning) or a wakefulness period. Speakers were placed to the left and right of the child, and the phrases were played from either direction randomly in order to measure the children's head turning and average looking duration (and thus deduce their familiarity with the rules). During the test phase (after 24 hours), children heard both phrases from the training phase but also new phrases (not heard at training). The new phrases had the same rule as the phrases heard at training.

In order for both information about specific individual items to be remembered and grammatical regularities to be generalised (as evidenced by a retest 24 hours following initial exposure), a nap had to occur within five hours and twenty-seven minutes (on average) after initial exposure. Infants who did not nap within approximately 4 hours after initial exposure did not show evidence of remembering specific individual items nor of having extracted grammatical regularities. This suggests that (for infants at least) a period of sleep should occur soon after (approximately 4 hours) following initial exposure to information that should be consolidated in its exact form or from which regularities should be extracted.

Conflicting findings come from a study by Werchan and Gomez (2013) with two and a half year old children. Children learnt new words (new names which belonged to three new categories called *'dax', 'tiv'* or *'fep'*). There were three objects per category (which the children were told were new toys). Within each category (e.g. *'dax'*), all three objects shared the same shape but varied in perceptual elements and colour. These objects (which were created using art materials) were photographed on different backgrounds (differing again in colour and texture). During training, the children saw all three objects belonging to a new category, while the experimenter named them (e.g. "Look at the '*dax*'!"). A distractor object was also presented but not named (so that participants would not rely on familiarity with objects alone to guide decisions at test).

The toddlers' ability to abstract and generalize the newly learnt words (i.e. the fact that the word was constant across the objects of the same shape but of different colour and background) was tested both immediately and following a consolidation delay, during which the toddlers either had a nap or stayed awake. The toddlers completed a four alternatives forced choice recognition task (choosing between an untrained new object belonging to a learnt object category such as a *'dax'*, a distractor (for that category), an unseen object or a familiar toy unseen at training). Participants could correctly name a new exemplar as a *'dax'* (for example), if they relied on the shape (which was shared and therefore generalizable) but not the variable background and colour. During this task, participants were asked to pick the object which corresponded to the new word heard through a recording (e.g. "Which one is a *'dax*?").

The children showed no evidence of generalisation ability (participants performed at chance at generalising learnt category names such as *'dax'* to untrained exemplars) when tested immediately following learning. A four-hour consolidation delay significantly improved generalisation ability (compared to the immediate test) only in infants who stayed awake (improving to above chance) but not in those who napped (whose performance remained at chance). The background and the colour of the object were 'irrelevant' details (in comparison to the 'relevant' words) which Werchan and Gomez (2013) argue may have been strengthened by sleep, therefore impairing the abstraction (and generalisation) of the word-object mapping. Such conflicting results may also be explained by the idea that forgetting of specific individual elements (such as the words themselves) could be necessary for some aspects of the generalisation process (e.g. Vlach, Ankowski, & Sandhofer, 2012).

To summarise, the above studies suggest that in infants a period of sleep-associated consolidation is required for generalisation to emerge and is not seen immediately. Wake-associated consolidation in infants benefits memories for exact individual items (Gomez et al., 2006; Hupbach et al., 2009) and in toddlers in some occasions it also benefits generalisation (Werchan & Gomez, 2013). In adults, generalisation may be evident immediately following training but declines with time (during the day). Only a period of sleep (and not the equivalent time awake) restores this generalisation ability (e.g. Fenn et al., 2003, 2013; Tamminen et al., 2013). Sleep is more likely to benefit generalisation ability when participants are exposed to a larger number of exemplars during training (Fenn et al., 2013) or when the trained exemplars have increased variability (e.g. Gomez et al., 2006). These findings suggest that sleep restores mappings of previously extracted regularities (representations for which are lost following training during the day) such that they can be applied to new exemplars.

1.2.1.3 Schema Formation: Relational Memory

Relational memory can be defined as the ability to create new representations based on the abstract relationships between separately learnt (existing) individual items, such that inferences based on these relations can be drawn (e.g. Cohen & Eichenbaum, 1993).

The formation of associations between distinct (but co-occurring) items into an integrated schema allows the preservation of the coherence of the existing knowledge even once new knowledge is added. According to relational theory, all associations which are encountered are connected into relational frameworks, a process which is mediated by the hippocampus (e.g. Cohen & Eichenbaum, 1993; Eichenbaum, 2004) to allow for inferences to be made based on these associations, even in novel contexts (Eichenbaum, 2000). Relational memory is considered distinct from item memory, which is the memory for the individual items in their original form without the formation of associations.

In adults, a task requiring relational memory provided evidence for the benefit of sleep for inferential ability, i.e. the ability to link together separate pieces of information (Ellenbogen, Hu, Payne, Titone & Walker, 2007). During training, participants learnt five related pairs (e.g. from '*A>B'* to '*E>F'*) during a hierarchical inferential learning task. The pairs had a hidden structure ('*A>B>C>D>E>F'*), the understanding of which was tested either after 20 minutes or after 12 hours (of either overnight sleep or day-time awake). Participants had a good knowledge of the specific pairs immediately (20 minutes) after initial exposure and the level of the knowledge of specific pairs did not change twelve or twenty-four hours after initial exposure. Inference ability (e.g. '*B>E'*) was at chance immediately (20 minutes) after initial exposure and improved significantly twelve and twenty-four hours after initial exposure. This suggests that inference ability unlike memory for individual specific items does not develop immediately and requires a period of consolidation to emerge. Moreover, participants who slept and those who stayed awake during the twelve-hour delay showed similar performance for inference pairs which were less distant. For pairs which were more distant (e.g. the pair '*B-E'* is two letters apart and therefore this relation is more distant compared to the pair '*B-D'* which is one letter apart), there was an additional benefit of sleep to inference (whether the 12 hour period of sleep occurred immediately i.e. a 12-hour delayed test or after a 12 hour day awake, i.e. a 24-hour delayed test). This possibly suggests that sleep (over wake) particularly benefits the relational memory of more difficult or weakest (not as easily observed or accessed) relations. Participants in the sleep group who showed better inference performance for the distant pairs did not show conscious awareness of this relational memory, suggesting that the benefits of sleep to relational memory (and inferential ability) may not require for information to be brought to conscious awareness in order for it to be accessed.

The idea that sleep benefits relational memory has also been shown in a study using a nap-paradigm, which required participants to complete an associative inference task (Lau, Tucker & Fishbein, 2010). During training, participants saw black and white photographs of human faces paired together with household objects (e.g. *'a cup'*). Each object was the relational element as it was paired together with two different faces. Immediately following training participants completed a four-alternatives forced choice recognition task assessing direct associative memory (of face-object pairs). On each trial, participants saw a photograph of a face and had to decide which one of the four presented objects was the correct pair (the remaining three incorrect objects were seen paired with different faces during training). After a consolidation delay (during which participants either napped or stayed awake) participants completed the same task. Relational memory (of faces which were paired with the same object) was also assessed in a forced-choice retention task. In this task participants saw a face and had to decide which one of the four presented faces was its correct pair based on the shared object (e.g. if the face was paired with a *'cup'* during training, its correct pair would be the face which was also paired with a *'cup'* during training).

Both associative memory (for face-object pairs) and relational memory (for face-face pairs) were significantly better in participants who napped than in those who stayed awake during the consolidation delay between training and test. Moreover, participants who experienced greater slow wave sleep (SWS) also showed better relational memory. Immediately following learning participants' ability to match faces with objects (i.e. associative memory for specific items) was on average 89% correct, reaching ceiling levels for the majority of participants. These findings suggest that even when immediate performance for individual items is near ceiling (compared to lower (i.e. high but not ceiling level) knowledge of individual items and poor (or at chance) immediate relational memory or overlapping rule knowledge, as with previous findings, e.g. Ellenbogen et al., 2007; Fischer et al., 2006), sleep (and particularly SWS) aids the reorganisation of memories in such a way that common relations between individual items learnt separately may be formed.

Similar performance was seen in the language domain (Cai, Mednick, Harrison, Kanady & Mednick, 2009) when relational memory was assessed using the remote associates test (RAT). Participants were required to identify a related link (which was a particular word such as *'sweet'*) between seemingly un-associated words (e.g. *'cookies', 'sixteen', 'heart'* – i.e. *'cookies'* are sweet, *'sweet sixteen', 'sweetheart'*). After a REM (rapid eye movement)-rich nap (as opposed to NREM (non rapid eye movement) or no-nap quiet rest) the ability to find these associations was improved. Participants in the quiet-rest, REM-rich nap and NREM nap performed similarly on tasks testing memory for specific items. Therefore, REM-rich sleep is beneficial specifically for activating associations between items (rather than just improving memories in general). These benefits of REM sleep for relational memory may occur due to the increased hippocampal acetylcholine together with decreased neocortical acetylcholine and norepinephrine. These properties (which accompany REM sleep) facilitate the reorganisation of semantic and associative networks to allow for the enriching of these networks through relational memory.

In another nap study (Lau, Alger and Fishbein, 2011), participants learned English translations for Chinese characters with semantic regularities (expressed through shared radicals, which were represented on the left hand side of the Chinese characters) and at test generalized these radicals to unseen characters, as well as explicitly saying what these radicals signified (e.g. the words *'sister', 'mother'* and *'maid'* shared the same radical based on their semantic category). Participants went to sleep or stayed awake (watching a video) either immediately following training, or after an additional 90 minute delay (during which wake group watched a second video).

Both participants who napped and those who stayed awake during the immediate consolidation delay (between training and test) showed similar recognition memory for individual specific pairs (i.e. the ability to correctly match trained Chinese characters with their English meanings). Participants in the delayed wake group had better recognition memory for trained characters than those in the delayed nap group. Participants who napped compared to those who stayed awake showed significantly better relational memory (i.e. ability to correctly match *untrained* Chinese characters with their English meanings). Participants who napped compared to those who stayed awake also showed significantly better explicit awareness of the trained Chinese characters' meanings (i.e. were better at explicitly naming the Chinese characters' meanings). These findings suggest that sleep (even as short as a nap) is beneficial for the formation of relational memories (i.e. links between) separately learnt specific individual semantic concepts and the formation of explicit awareness for these concepts but not necessarily for the consolidation of the specific individual items on their own.

In summary, these data show that inferential ability in adults does not always emerge immediately (even with a high level of knowledge of the individual items) but requires a period of at least 12 hours of overnight sleep (but not the equivalent time awake) to emerge (Ellenbogen et al., 2007). Even in studies where all participants reached near ceiling levels of accuracy for individual items by the end of training and the sleep period was shortened to a nap, participants in the sleep group (compared to those who spent the equivalent time awake) were still found to show better relational memory both for non-linguistic associations (Lau et al., 2010) and for linguistic associations (Lau et al., 2011). Moreover, despite better relational memory as well as increased conscious awareness of the relations between items, participants in the sleep group did not have better recognition memory of individual items than those in the wake group following a period of consolidation (Lau et al., 2011). These findings suggest that a delay between training and test involving sleep provides an additional benefit over the equivalent time awake for relational memory, but that relational memory may be consolidated and retrieved through different processes than memory for individual items.

1.3 Declarative and procedural learning and consolidation

The declarative/procedural (DP) model (e.g. Ullman 2001; 2005; 2015) is based on evidence suggesting the existence of two separate domain *general* memory systems, namely the declarative system and the procedural system. The DP model suggests that on the basis of existing knowledge about the two domain *general* memory systems, predictions can be made as to how the learning, consolidation and use of *language* might be linked to these domain general memory systems. A brief summary of the DP model is provided next, in order to inform such predictions about language.

The acquisition, consolidation and use of information relating to both facts (i.e. semantic information) and events (i.e. episodic information) is subserved by the domain-general declarative system. In addition, patients with amnesia who have damage to the MTL often have difficulty acquiring new information of an arbitrary nature, suggesting that the MTL and the declarative system might subserve the acquisition of arbitrary information (e.g. Henke, 2010). Learning and consolidation within the declarative memory system is mainly subserved by the hippocampal structures, as well as the medial temporal lobe (MTL) (e.g. Ullman, 2001, 2004). With time, information in the MTL develops links with neocortical representations (e.g. McClelland et al., 1995). Information which is subserved by the domain-general procedural system is typically implicit in nature, and includes the learning of regularities, sequences and motor skills (e.g. Henke, 2010). Learning and consolidation with the procedural memory system is mainly subserved by the frontal and basal ganglia areas (e.g. Ullman, 2001, 2004). It is important to also bear in mind that there is some interaction across the two memory systems, consequently the acquisition and retrieval within one memory system may prevent information from being acquired and retrieved in the other memory system (Ullman, 2004).

The DP model (e.g. Ullman, 2001, 2015) suggests that acquisition within the procedural system requires a longer time period than acquisition within the declarative system, as learning within the procedural system requires additional exposures, i.e. additional practice. This is important for predictions as to how the two domain-general memory systems *might* be related to the learning, use and consolidation of language. The DP model predicts that the learning of phonological forms and form-meaning mappings might be linked to the declarative system for both first and second language learners. The learning of aspects of grammar *might* rely on the procedural system. However, given that (as mentioned above) learning within the procedural system might require additional practice and increased fluency (or proficiency), the learning of grammar in a second language might initially be related to the procedural system.

Indeed, previous studies have provided support for the predictions of the DP model (Ullman, 2001) that aspects of language might be related to the two domain-general systems. For instance, in the declarative domain, Breitenstein et al. (2005) investigated the neural substrates of novel word acquisition. Participants learned novel, arbitrary word-picture pairs (e.g. "bini – book") while in the fMRI scanner. In the test measuring novel word recognition (outside of the scanner), participants heard a real German word, followed by a novel trained word. Participants' task was to decide whether the translation was correct or incorrect.

It was found that the learning of new, arbitrary lexical items was related to increased activity in the left hippocampus. The findings suggest that the acquisition of arbitrary form-meaning mappings in language is linked to the hippocampal system, just as the domain-general declarative system is subserved in part by the hippocampus (e.g. Ullman, 2001). This suggests that the learning of arbitrary mappings may be linked to the declarative system. Moreover, as lexical ability increased, the hippocampus (and also the fusiform gyrus associated with crossmodal mapping learning) became deactivated. This suggests that arbitrary semantic mappings may initially be learned via the hippocampus, but with stabilization may rely more on the necortical system instead (e.g. Squire & Zola Morgan, 1991).

As suggested earlier in this Chapter, Ullman's (2001) declarative-procedural model predicts that the acquisition of grammar *might* be linked to the domain-general procedural memory system. In support of this, in the procedural domain, another study suggests that the link between grammatical regularity consolidation and the procedural systems depends on individual procedural consolidation ability (Brill-Schuetz & Morgan-Short, 2014). Over a period of four days participants learned an artificial language, which complied with a subject-object-verb sentence order (using novel nouns of a masculine or feminine gender such as "pleck", and novel verbs such as "kiln", as well as gender adjectives, gender determiners and adverbs). The participants' task was to engage in a "game", by listening to sentences in the artificial language and then either making a move on the computerized game board to carry out the sentence instruction, or to provide a verbal description of the sentence. For example, if participants heard the sentence "*Blom nim lu neep neime li praz."* it meant "The capture blom-piece switches with the square neep-piece." (blom and neep being game pieces). Participants learned 20 new sentences. Only participants in the explicit (but not the implicit) learning condition were told the rules of the language and given written sentence examples.

Participants' memory was tested using a grammaticality judgement task. This involved deciding whether untrained sentences were correct or incorrect in terms of grammatical word order (and sentences with the incorrect word order were included by swapping word classes in the sentences, e.g. putting a noun in the place of a verb). To measure procedural memory, an alternating serial reaction task was used (an implicit sequence learning task). In this task, circles seen on screen corresponded to keyboard keys. Participants were instructed to watch out for a circle on screen being coloured in (black), and to immediately press the corresponding key. There was an underlying pattern of the order in which the circles were coloured in (e.g. circle 1 – random circle – circle 2 – random circle – circle 3 – random circle – circle 4 – random circle). The weather prediction task was also used to measure procedural memory (predicting rain/sun on the basis of geometrical patterns on cards).

It was found that participants with higher procedural memory (in comparison to those with lower procedural memory) had higher grammatical task performance, especially if grammatical information was learned implicitly. (In this case the term "implicit" is used to refer to learning without instruction on grammatical rules.) The findings suggest that whether grammatical regularities may be linked to the procedural system also depends on the learner's procedural memory ability.

There are studies which have included both declarative and procedural tasks for comparison. For example, in Ettlinger, Bradlow and Wong's (2012) study participants were exposed to new word-picture mappings referring to animals. The novel words were created using stems, 1 diminutive-marking prefix and 1 plural-marking suffix. The input consisted of both simple and complex morphophonological patterns. The simple pattern could *always* be derived from the context, in this study this meant that in combining stems with suffixes (or stems with prefixes), *no* vowels had to be altered. Vice versa, the complex pattern could *not* always be derived from the context, and was rather analogy-based, in this case meaning that in combining stems with suffixes, vowels *were* altered.

Participants' generalisation ability was tested by presenting untrained pictures. Participants were required to choose the corresponding form for the picture. Participants' procedural memory was measured using a tower of London task. Participants used a cursor to move balls on screen into a target arrangement sequence (and participants were tested on the same sequences in order to measure changes in ability between test and retest). Declarative memory was measured using a visual-auditory task. In this task participants acquired associations between abstract symbols and common English words.

All participants who were able to generalize the complex morphophonological pattern to untrained words, were also able to generalize the simple morphophonological pattern in the language task. However, there were participants showing generalization of only the simple (but not the complex pattern). It was more difficult to learn the complex pattern than the simple pattern. Moreover, procedural memory as measured by the tower of London task (but not declarative memory) correlated with simple patterns' generalisation. Declarative memory as measured by the visual-audiotory task (more so than procedural) correlated with complex patterns' generalisation. Overall the findings suggest that learning simple grammatical patterns can be linked to the procedural system, whereas learning complex analogy-based patterns can be linked to the declarative system.

Moreover, Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter and Wong, (2014) investigated whether the implicit grammatical acquisition in a second language can be linked to the declarative or to the procedural memory system. To measure declarative memory both verbal and nonverbal tasks were used. The verbal task required learning novel words paired with their English word translations. To assess nonverbal declarative ability, the continuous visual memory task was used. Participants saw abstract images on screen and later were tested on their memory by indicating whether they had seen each item previously ('old') in the task or not ('new'). Procedural memory was assessed using the weather prediction task and the tower of London task (as in the previous study). The same artificial language was used as described in Brill-Schuetz and Morgan-Short (2014), the learning of which was similarly assessed using a grammaticality judgement task.

The initial stage of the ability to learn the grammar was related to solely the declarative system. This was no longer the case at the later stages of learning, as then grammatical ability was related to solely the procedural system (as measured by both the weather prediction task and the tower of London task). Potentially this is because learning is faster within the declarative system, but requires more time (with increased exposure) within the procedural system.

Further neuroimaging evidence supports the idea that language aspects may be linked to the declarative and procedural memory systems (Lieberman, Chang, Chiao, Bookheimer & Knowlton, 2004). This study was based on the premise that the declarative system is subserved by the hippocampus and the medial temporal lobe (MTL), whilst the procedural system is subserved by basal ganglia, and in particular the caudate nucleus (e.g. Ullman, 2001). In the study (Lieberman et al., 2004), participants learned novel grammatical strings. Half of the grammatical strings were of high chunk strength (meaning the bigrams and trigrams within the strings appeared frequently in the input). Half of the items were of low chunk strength. Participants firstly saw novel grammatical strings, which they had to repeat. Next, participants were informed that the strings complied to underlying rules. The participants' task was to respond (by guessing if they had not acquired the rules) whether each string was consistent with the underlying rules, while participants' neural activation was measured.

It was found that the hippocampus was activated when higher strength chunks were retrieved (regardless of their grammaticality). The MTL was also activated when high strength nongrammatical chunks were retrieved. Conversely, the caudate nucleus showed greater activation for grammatical than ungrammatical items. Moreover, as caudate activity was increased, the hippocampus became deactivated, suggesting that there is competition between the two neural systems. The findings suggest that the caudate nucleus is linked to implicit grammatical rule retrieval, whereas the hippocampus (and the MTL) are linked to the retrieval of trained item chunks. Given that the caudate nucleus in general memory mechanisms subserves the procedural system, and the hippocampus and the MTL subserve the declarative system, the evidence suggests that some aspects of grammar may be linked to the procedural system, whereas trained item retrieval is linked to the declarative system.

1.3.1 The role of sleep in procedural and declarative memory consolidation

The previous section has discussed the possibility of links between procedural and declarative memory systems and aspects of language, regardless of sleep versus wake-dependent consolidation. There is also evidence suggesting that the consolidation of both declarative and procedural memories is differentially affected by wake compared to sleep-dependent consolidation. More specifically, sleep may play a beneficial role in the consolidation of both declarative and procedural memories. According to the "dual process hypothesis" (e.g. Plihal & Born, 1997), based on previous evidence, SWS (as opposed to REM sleep) may be beneficial for the consolidation of the more hippocampally-dependent declarative memories (e.g. Rasch, Buchel, Gais & Born, 2007). Meanwhile, REM sleep (as opposed to SWS) may be beneficial for the consolidation of the less hipocampally-mediated procedural memories (e.g. Wagner, Hallschmid, Verleger& Born, 2003). Further evidence suggests that this distinction may not be as clear cut as initially suggested. For instance, in one study, the amount of REM sleep was found to be related to sleep gains on the procedural finger-tapping task (Fischer, Hallschmid, Elsner & Born, 2002), but this correlation was not replicated in the follow up study (Fischer, Nitschke, Melchert, Erdmann & Born, 2005).

Moreover, two further studies have found that stage 2 sleep and sleep spindles amount (as opposed to REM sleep amount) may be related to sleep gains on the procedural finger-tapping task (Nishida & Walker, 2007; Walker, Brakefield, Morgan, Hobson & Stickgold, 2002). There are also studies which have found a benefit of SWS (as opposed to REM sleep) for the consolidation of procedural memories (e.g. Antony, Gobel, O'Hare, Reber & Paller, 2012). Hence, this distinction (of REM sleep mediating procedural memory and SWS mediating declarative memory consolidation) appears to be an oversimplification, as evidence is somewhat contradictory. One *possible* explanation for this discrepancy in findings is that the distinction between procedural memory and declarative memory is not entirely clearcut, and procedural memories interact with declarative memories, and thus have an underlying hippocampal component on which sleep can act (Albouy, King, Maquet & Doyon, 2013).

Moreover, there is evidence to suggest that at least in the procedural domain, whether sleep has any effect on consolidation at all, depends on the initial level of acquisition. Wilhem et al., (2012) investigated whether initial level of performance affects sleep-dependent consolidation of motor memories. The amount of pre-sleep training was varied in both children (aged 4–6 years) and adults using a coarse motor sequence task (i.e. the button-box task). The button box had eight coloured buttons, which lit up in an eight part sequence. The participants' task was to press the button as it lit up, as quickly as possible. Sleep benefited the consolidation of procedural motor memory only when the initial level of acquisition was intermediate. One possible explanation of this finding is that the initial level of acquisition is related to the amount of hippocampal activation. For example, children with lower performance may have had lower hippocampal activation (Thomas, Hunt, Vizueta, Sommer, Durston, Yang & Worden, 2004) and the correct amount of hippocampal activation is crucial for sleep-dependent consolidation. In another study described later in this review (Albouy et al., 2008), post-sleep improvements in motor sequence learning were similarly related to hippocampal activity.

Moreover, while initially the fronto-striatal networks were activated in the motor sequence learning, with time they become deactivated, as hippocampal activity increased. This might be because the cortico-striatal areas typically associated with procedural memory systems (e.g. Ullman, 2001) might underlie specifically implicit learning, whereas the hippocampal systems might underlie specifically explicit learning. With time, explicit awareness of implicit information might be gained (e.g. Albouy et al., 2008; Wilhem et al., 2012).

The correlations mentioned above between REM sleep and procedural tasks (e.g. Fischer, Hallschmid, Elsner & Born, 2002; Wagner, Hallschmid, Verleger& Born, 2003) have been found when the tasks were used in isolation. There is evidence from Barsky, Tucker and Stickgold (2015) suggesting that this correlation (between REM sleep and procedural task performance) is not observed when an additional interference task is also used. Participants completed a weather prediction task (WPT). In this task, participants were exposed to a set of cards. Each card depicted a drawing of an object, selected from different categories (e.g. animal, vehicle, etc.). Participants were unaware that each card had a probability of predicting the Sun or Rain. After a 90 minute period (which consisted of either a nap or wakefulness), half of the participants were retrained on a second, novel set of cards which had the same underlying probabilities, in order to create interference with the first set. Without interference, it was found that the 90 minute nap (but not the equivalent time awake) resulted in improved WPT performance (i.e. ability to predict Sun or Rain). This improvement was related to REM sleep amount. As a result of the additional interference, the post-nap improvement was no longer observed (the performance of the wake group was similar regardless of interference). Crucially, the interference reduced the specific sleep-dependent *gains*, as opposed to resulting in an overall loss of WPT memory.

Similar results have been found in the declarative domain. Deliens, Leproult, Neu and Peigneux (2013) investigated whether sleep protects memories from retroactive interference. Participants took part in the experiment after a night of sleep, the overall length of which was reduced by 2 hours. This led to increased SWS sleep in the experimental nap. Participants learned their first set of unrelated, novel word pairs. Next, participants either had a morning nap lasting 45 minutes, or remained awake. Subsequently, participants learned a second set of interfering word pairs (in the interfering list, half of the pairs, had the same first word as the pairs from the first set, in order to cause interference), followed by delayed recall. Recall of interfering pairs was positively correlated with slow oscillation power. Recall of the first set was less affected by interference when SWS was present in the nap (than when SWS was not present). This suggests that SWS may protect declarative memories from retroactive interference, and this protective effect may be mediated by slow oscillation power.

To summarise, both Barsky, Tucker and Stickgold (2015) and Deliens, Leproult, Neu and Peigneux (2013) found that when interference is within one memory system, wake-dependent consolidation is not affected. Sleep may have a protective effect against interference within the declarative memory system. Within the procedural memory system, interference leads to the loss of post-sleep gains that would be observed without interference.

1.3.2 Interference and interaction across modalities

Interference arises not only from the acquisition of within-memory system information, but also from the acquisition of across memory systems information. On the basis of evidence of classification learning experiments with fMRI, Poldrack et al., (2001) initially suggested that there may be *competition* between brain systems in the acquisition of cross-modality information. In this study, two versions of the weather prediction task were used. One group of participants completed the non-declarative version of the weather prediction task, which was feedback based. Another group of participants completed the declarative version of the weather prediction task, which relied on paired-associate learning. A third group was tested on the feedback-based task using event-related fMRI (to look at changes in activation across learning).

The nondeclarative task showed greater activation in the striatal areas (i.e. the basal ganglia) and cortical areas, but lower activation of the medial temporal lobe (MTL), vice versa for the declarative task. Moreover, the activation of the right caudate region and a left MTL region were found be negatively related. To begin with, while the MTL was engaged, there was no caudate engagement. As the caudate became engaged, the MTL disengaged and did not become engaged again. Meanwhile, the engagement of the caudate altered between decreases and increases. These results have shown that the learning of declarative information is predominantly subserved by the MTL, and the learning of nondeclarative information is predominantly subserved by striatal areas. Secondly, the negative relationship of the MTL and the striatal areas, and the observed changes in engagement, point to the existence of *competition* between brain systems (more specifically between the MTL and striatal areas), in the acquisition of declarative and nondeclarative information.

There is evidence to suggest that this competition between brain systems may lead to interference across modality, when acquiring both declarative and procedural information. Brown and Robertson (2007) investigated whether the declarative component of a task inhibits the offline consolidation of the procedural (i.e. motor skill) memory of the same task, looking at wake-dependent consolidation only. Participants were trained and tested on the SRTT task (with both a motor-skill procedural and a sequence learning declarative component). Immediately after this, 1 group of participants learned a word list (to engage the declarative memory). A different group of participants did a nondeclarative but cognitively demanding vowel-counting task instead (i.e. counting the number of vowels in a nonsense letter string). After 12 hours, all participants were re-tested on both the procedural and the declarative component of the SRTT.

Engaging the declarative system by learning the word list significantly decreased performance on the declarative component of the SRTT (i.e. free recall). When the SRTT was followed by the declarative word list (but not by the non-declarative vowel-counting task), improvements on the SRTT motor component were observed after a consolidation delay. Hence, the declarative component of the SRTT task may have interfered with the learning of the procedural component of the same task. The engagement of the declarative system (using the word list) disrupted the declarative component of the SRTT. As a result, the procedural component was enhanced. The procedural and declarative memory systems interact and compete with each other. As a result of this, the disengagement of one component of a task, may improve the consolidation of another component of the same task.

There is also evidence to suggest that sleep plays a role in this interference. Brown and Robertson (2007) investigated whether there was interference across memory modalities (from declarative to procedural), and how this interference was affected by a consolidation delay. In one scenario, the procedural motor task (SRTT) was learned first, followed by a declarative, word learning task second. (Another group of participants completed the vowel-counting task instead of the word-learning task).

In another scenario, the task order was swapped. In other words, the declarative task was learned first, followed by the procedural task. (Another group of participants completed the random reaction time task instead of the SRTT in this scenario). After 12 hours (consisting of either daytime wakefulness or overnight sleep), participants were retested on the tasks. Those who were tested on the procedural motor SRTT were also tested on the declarative component of the SRTT, i.e. the sequence recall. An additional control group was tested either 30 minutes or 4 hours following acquisition, instead of 12 hours.

In terms of wake-dependent consolidation, the second task disrupted the consolidation of the first task, regardless of the memory type (i.e. regardless whether the first task was declarative or procedural). There was no disruption over sleep-dependent consolidation. Moreover, the decrease in motor skill correlated with word recall, and the decline in the declarative SRTT recall also correlated with motor skill. These correlations suggest that procedural and declarative memories are not completely distinct systems, but have reciprocal interactions. On the basis of these interactions, and as a result of shared neural circuits, interference arises. The sleep state may alter the connectivity between neural circuits, disconnecting the procedural and declarative systems, allowing them to operate distinctly.

In order to investigate how this interference arises at the neural level, Cohen & Robertson (2011) used TMS. Participants learned and recalled a word list, followed by motor skill learning. At the end of this learning session all participants' performance was similar on both tasks. Participants were then separated into three groups: sham stimulation, TMS to primary motor cortex (M1) and TMS to right dorsolateral prefrontal cortex (DLPFC). Both the M1 and DLPFC were suggested to play a role in executive control, and hence these areas were chosen because they might be actively generating interference across memories. Following 12 hours, participants were retested on the word recall and the motor skill task.

Importantly, participants from all 3 groups had similar levels on the motor skill retest. Word recall was found to have significantly decreased after both sham and M1 stimulation. Moreover, this decrease correlated with initial skill on motor skill learning. Crucially, after DLPFC stimulation, there was significantly less decrease in word recall. To distinguish whether the DLPFC stimulation affected performance by enhancing word recall in isolation, or in interaction with the motor skill learning task, the motor skill *learning* was replaced with motor skill performance. There was now a significant decrease in recall after DLPFC stimulation. This suggests that stimulation affected the interaction of the tasks, as opposed to any of the tasks in isolation. Additionally, to support this, the correlation between the word recall and motor learning tasks was reduced when TMS was applied to the DLPFC.

Next, the task order was reversed, i.e. the participants completed the motor skill learning before the word recall. After 12 hours, sham stimulation and DLPFC stimulation both resulted in significant motor skill decrease, which correlated with word recall. M1 stimulation significantly increased motor memory, with no interference from word-learning, although word recall changes were not different from the other 2 groups. When the word recall task was replaced with the vowel counting task (nondeclarative in nature), there was a trend towards decrease in motor skill after M1 stimulation. The correlation between tasks was also reduced.

These findings suggest that memory interference results from communication between different neural areas. This communication between neural areas can be disrupted using TMS, hence decreasing the amount of resulting interference between tasks.

Keisler and Shadmehr (2010) based their interference tasks on those used in Brown & Robertson (2007), to investigate whether the more rapid motor acquisition is subserved by the declarative system (unlike the slower motor acquisition). The premise of this study was that the speed of acquisition is a key difference between the procedural and declarative systems. Information is acquired more rapidly within the declarative system (and as a result is also more rapidly forgotten), than within the procedural system. It was hypothesized that the more rapid motor acquisition is subserved by the interaction of the procedural and declarative systems, but as a result the more rapid process is more sensitive to errors, and more prone to interference from the subsequent acquisition of declarative information (compared to the slow process).

In this study, participants completed the single-target task, requiring motor adaptation. In this task, participants used a cursor on screen, to move a robot's arm in order to reach a target. Two force fields with opposing directions (in order to cause competition) were used. Exposure to field A was long, leading to slow motor acquisition. Exposure to field B was brief, leading to rapid motor acquisition. The interference tasks in this study included a declarative memory task (word-pairs learning, tested using cued recall), and a non-memory task (i.e. a non-declarative task, saying aloud the number of vowels in a letter string).

Declarative information did not interfere with the slower motor consolidation following longer exposure to field A, but interfered with the motor consolidation followed brief exposure to field B. When a 6 hour consolidation delay was included prior to the test, the declarative information interfered with the motor memory of the second, brief field B. Hence, the brief exposure resulted in a more rapid acquisition, which was also more susceptible to forgetting, and was subserved by the declarative system. Following a consolidation delay, the slower motor consolidation was boosted by the more rapid and declarative memory being disrupted. This suggests inhibition between the declarative and nondeclarative systems. Hence, the findings of this study were consistent with Brown and Robertson's (2007) results that a procedural task (expectedly striatally-dependent) and a declarative task (expectedly hippocampally dependent) interfere with one another.

Indeed, there is support for this hypothesis at the neural level, provided by Albouy et al., (2015). The 55 participants were firstly explicitly trained on a motor finger-tapping sequence (4-1-3-2-4) using a keyboard. Next, participants underwent 3 sequence-tapping practice sessions in the fMRI scanner. Participants were then separated in to one of four groups: allocentric nap, allocentric wake, egocentric nap or egocentric wake. Then, the hand and the keyboard were both turned upside down. The allocentric (i.e. spatial) groups were asked to type 1-4-2-3-1, meaning that the finger movement had changed, while spatially the location was the same as the prior training and the upcoming retest. The egocentric (i.e. motor) groups were asked to type 4-1-3-2-4, meaning that the finger movement remained the same, while spatially the location had changed. Participants' speed and accuracy was measured. Next, the consolidation delay (involving nap or wakefulness) lasted 90 minutes, before all participants were retested.

On the behavioural level, for the allocentric (i.e. spatial) groups, it was found that only participants from the allocentric-nap (importantly, not the allocentric wake) group showed improvement at the retest. The performance of participants from the egocentric (i.e. the motor) groups was maintained regardless of whether they had a nap or remained awake. There was some numerical (marginally significant) evidence of loss of motor memory following wake.

Neurally, there was evidence to suggest that allocentric-spatial information was subserved predominantly by the hippocampal (and cortical) areas. Meanwhile, egocentric-motor information was subserved by the striatal areas. Moreover, in the egocentric (i.e. motor) sleep group (but not egocentric wake group), the engagement of the hippocampus at learning was positively related to post-sleep performance. This suggests that the protective effect of sleep in motor memory maintenance is enhanced by hippocampal engagement at learning. Regardless of sleep, for the allocentric, i.e. spatial groups (associated with hippocampal activation), the more the striatum was engaged during learning, the poorer the post-delay performance was. This suggests that striatal engagement interfered with the hippocampal areas and hence with allocentric improvement post consolidation.

Overall, the findings suggest that motor sequence learning may engage not only the striatal areas (as typically expected), but also the hippocampal areas. Specifically, while the motor (i.e. involving finger movement) component of the motor sequence engages the striatal areas, the spatial component of the motor sequence (i.e. the layout of the keyboard) engages the hippocampus at learning. This engagement of the hippocampus enhances the protective effect of sleep on motor memory, i.e. sleep additionally protects motor memory from loss.

Hence, the above evidence suggests that when both a declarative and a procedural task are completed in one session, there is interference between the systems in the acquisition and also consolidation of the information.

Crucially, there is also evidence that further exposure to declarative and procedural information at *re-test* may also cause interference at retrieval, i.e. not only during the acquisition and consolidation phases. For instance, there is evidence from Albouy et al., (2008) to suggest that the interference between declarative and procedural information is not limited to the acquisition phase. They investigated neural activation of an implicit oculomotor sequence task not only during acquisition, but also at retrieval, over a period of 24 hours. In this task, participants were exposed to an eight-element sequence of coloured dot locations. Eye movement latencies were used as an indication of the participants' implicit acquisition of this sequence. For fast learners (i.e. with fast eye latencies, indicative of implicit sequence acquisition), the engagement of the hippocampus and striatal areas at learning was positively related to the extent of the overnight gains on the task. There was no such correlation in slow learners (i.e. with sloe eye latencies), as there was no evidence of overnight gains on the task

Crucially, when the fast learners were re-tested on the sequence twenty-hours following initial acquisition, the hippocampal and striatal areas were again engaged. This engagement corresponded to the observed gains on the task. Hence, this study did not only provide evidence for the engagement of hippocampal, as well as striatal areas during motor sequence *acquisition*, and the importance of this engagement for subsequent overnight gains in performance. It additionally showed that these areas are also engaged during the *retrieval* of the motor sequence. These findings suggest that the interference or interaction between the areas may be present not only during acquisition, but also at retrieval.

Building on these findings, Gagné and Cohen (2016) investigated the extent of the interaction between the declarative and procedural systems during retrieval (as opposed to during initial acquisition). Three tasks were used in this study. In the FOS task, participants completed the sequence 4-2-3-1-4, by opposing the corresponding finger (which was assigned a number 1-4) against the thumb. In the visual search task, participants were instructed to locate targets in pictures (e.g. "Locate three pigs in the Carnival picture"). In the visuospatial task (completed by the declarative group only), participants recalled spatial details about the original picture from memory. Two groups of participants took part in two morning sessions spaced 24 hours apart. On the first day, both the control and declarative groups completed the FOS task, followed by the visual search task. On the second day, both groups were tested on the FOS task. The declarative group additionally completed a visuospatial memory task prior to retrieval of the FOS task.

The control group showed improved performance on the FOS task, but the declarative group did not. Hence, retrieving spatial information prior to retrieving FOS task information prevented any FOS task improvement at retrieval. Given that the improved FOS performance was present in the control group, it suggests that the interference occurred at retrieval, as opposed to during acquisition. This is because the acquisition session was the same for both groups.

The reason why there was no interference from the visual search task at acquisition for the two groups may be because this task is not expected to involve the hippocampus or the striatal areas. These are the areas that subserve declarative and procedural motor learning, and hence the areas where interference would occur (Kastner & Ungerleider, 2000). Hence, the interference may have arisen as a result of both the spatial components of the FOS task and the visuospatial task being subserved by the hippocampus. This study has shown that interference arises not only during acquisition, but also during retrieval. The extent of the interference may depend on the extent to which components of the different tasks rely on the same neural areas.

1.4 Conclusions and thesis outline

In summary, firstly it has been shown in this review that sleep plays an active role in language learning and integration of newly learned words into the mental lexicon through stabilization and strengthening of the newly learnt individual items in their exact form. However, there are cases (which may be task specific) in which wake-associated consolidation may be sufficient for integration. The new dimension which this thesis will add is the distinction of systematicity and arbitrariness as part of language learning. This idea draws on the views presented within the connectionist framework. This framework suggests that arbitrary mappings are learned in their exact forms (due to lack of systematicity). Systematic mappings meanwhile, have shared regularities which can be abstracted and applied to new exemplars (a process termed generalisation). According to sleep-related consolidation models, sleep facilitates the process of rule abstraction and generalisation through reorganisation of memories during their offline replay. Previous evidence has defined this as schema formation based on abstraction of regularities and as evidenced by generalisation. Sleep has been shown to support these processes through restructuring newly acquired memories to form a new schema. However, as in order for generalisation to occur, forgetting of some individual items may be necessary, sleep may not always lead to improved generalisation (for example in cases where there are other, non-generalizable elements which sleep may selectively strengthen at the expense of strengthening the generalizable elements). Similar issues have also been addressed in literature on relational memory (i.e. cases where separately learned elements have a shared i.e. relational feature to be consolidated), which has also shown that sleep over the equivalent time awake benefits the formation of relational memory (for shared elements).

Remaining unanswered questions which this thesis is going to address are how adults acquire and consolidate new individual arbitrary and systematic form-meaning mappings, and additionally whether sleep provides any benefits over the equivalent time awake for the consolidation of these mappings. These questions will be addressed in Chapters 2-5 by training participants on an artificial language with varying levels of systematicity. More specifically, in Experiments 1, 3, 4, 5 and 6 the sleep group will be trained in the evening and tested in the morning, vice versa for the wake group. This will be done in order to compare sleep-dependent versus wake-dependent consolidation of arbitrary and systematic mappings behaviourally. In Experiment 2, overnight polysomnography data will be collected to relate sleep parameters (and SWS specifically) to the consolidation of individual arbitrary mappings.

Additionally, across the six Experiments factors which may affect the extraction of regularities contained within systematic mappings will be investigated. Specifically, Experiments 1-3 will investigate whether the degree of systematicity is the only crucial factor which may determine the extraction of regularities. In other words, it will be investigated whether the *more* systematic mappings are *always* extracted, regardless of the location of the systematicity (i.e. regardless of whether the *more* systematic mapping is at the beginning of the word, or at the ending of the word). Experiment 4 will investigate whether additional training boosts the extraction of the regularities and whether the regularities are also available immediately. Experiments 5 and 6 will investigate which factors other than systematicity may determine the extraction of regularities. Specifically, firstly it will be investigated whether there is a redundancy of multiple cues (i.e. if two regularities both map on to the same meaning, whether the extraction of one of those mappings is inhibited due to its redundancy). Secondly it will be investigated whether increased exemplar variability boosts the extraction of regularities contained within systematic mappings. The findings from the six Experiments collectively will provide suggestions as to the language factors which may affect language acquisition beyond sleep. To summarize, it is firstly of interest to investigate whether sleep plays a role in language consolidation processes, both for arbitrary mappings and for systematic mappings. Secondly, it is of interest to investigate factors relating to the structure of the language input which may affect the extraction of regularities contained within systematic mappings.

A separate and second issue which this review has raised is the dinstinction between declarative and procedural memory. Specifically, based on the evidence presented, a much more complex picture emerges than the simple distinction between declarative and procedural memory systems and how the two systems might be affected by sleep versus wake-dependent consolidation. Initially, evidence suggested that procedural memory may be facilitated by the REM stage of sleep (e.g. Wagner, Hallschmid, Verleger & Born, 2003) or stage 2 sleep (e.g. Walker et al., 2002), and show sleep-dependent gains. Declarative memory may be facilitated by SWS and show less forgetting following sleep (Rasch, Buchel, Gais & Born, 2007). However, this may not always be the case. For instance, whether sleep benefits procedural memory or not, may depend on the initial level of acquisition, whereby sleep may only benefit procedural memory if it was initially acquired to an intermediate level (Wilhelm et al., 2012).

The picture becomes more complicated when more than one task is acquired within one session. When the two tasks are both within modality, there is evidence to suggest that for procedural memory, the interference reduced sleep-dependent gains, but did not affect wake consolidation (Barsky, Tucker & Stickgold, 2015). For declarative memory, there is evidence to suggest that SWS may protect memory from retroactive interference (Deliens, Leproult, Neu & Peigneux, 2013). When two cross- modality tasks are acquired within one session, (i.e. a declarative and a procedural task), there is evidence to suggest that there is competition between neural areas at acquisition, leading to interference (Albouy et al., 2015; Brown & Robertson, 2007b; Cohen & Robertson, 2011; Keisler & Shadmehr, 2010; Poldrack et al., 2001). The acquisition of a second task interferes with the acquisition of the initial task across wakefulness, but not across sleep

(Brown & Robertson, 2007). Moreover, this interference is not limited to acquisition and consolidation of information, but also occurs during retrieval of information (Albouy et al., 2008, Gagne & Cohen, 2016). These points will be addressed in this thesis in Chapter 6 as an exploratory investigation. Specifically, across all six of our Experiments in addition to the language tasks we are going to administer a procedural memory and a declarative memory task, both in session one and session two. All of the tasks (i.e. the language tasks, the procedural task and the declarative task) will be administered in one training session (i.e. session one). Following a consolidation delay, the procedural memory, the declarative memory and the language memory will all be tested in one session as well (i.e. session two). This means that interference may arise at both acquisition (in session one) and at retrieval (in session two). Consequently, it will be of interest whether we can replicate the previous sleep effects found for procedural memory (e.g. Walker et al., 2002; 2003) and declarative memory (e.g. Rasch et al., 2007), despite the likelihood of intereference.

In addition, on the basis of the presented studies, the declarative system might be linked to individual phonological forms and arbitrary form-meaning mappings, because both rely on the hippocampus and the MTL for encoding and consolidation. It is reasonable to suggest that the declarative *can* initially also process grammatical regularities (e.g. as shown by Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter and Wong, 2014). Despite this capacity, it cannot be assumed that the declarative system would be linked regularity extraction. The procedural system *might* be linked to implicit grammatical regularity learning and extraction. This is because both rely on frontal and basal ganglia neural areas (and in particular the caudate nucleus and anterior putamen). It might be predicted that initially, grammatical regularities might be linked to the declarative system, but as extraction ability improves, they might develop links with the procedural system. Even if that is the case, it does *not* mean that grammatical regularities will no longer be linked to the declarative system, once they have developed links with the procedural system, given that the two memory systems may interact. These concepts will also be addressed in Chapter 6 as part of the exploratory investigation, with the aim to explore whether the memory for arbitrary language mappings and systematic language mappings can be linked to the domain-general declarative or procedural memory systems.

Chapter 2: Sleep-related consolidation of new form-meaning regularities

2.1 Introduction

The aim of the first Experiment in this thesis was to investigate the influence of sleep-dependent memory consolidation on learning new (phonological) form-meaning mappings (as found in natural languages and re-created using an artificial language) in adults. Particularly, the differences in learning and consolidation of individual specific items were compared to the extraction of overarching regularities.

As discussed in Chapter 1, previous evidence has shown that sleep is beneficial for the consolidation of new phonological forms (i.e. spoken words) in adults. For example, Dumay and Gaskell (2007) have found that a period of sleep firstly improves the stabilization of individual specific items. Secondly, a period of sleep benefits the integration of newly learnt spoken words into the mental lexicon, in this case by enhancing existing connections between spoken word forms. The findings that a period of sleep (compared to the equivalent time awake) between training and test boosts consolidation of new phonological forms have also been reported using a similar paradigm (spoken word learning including 12 hour and 24 hour and also 1 week retests) and similar tasks in adults (e.g. Tamminen & Gaskell, 2008) and in children (e.g. Brown, Weighall, Henderson & Gaskell, 2012; Henderson, Weighall, Brown & Gaskell, 2012; Williams & Horst, 2014).

As well as findings showing that sleep benefits the consolidation of specific individual items (in the phonological form), previous findings have also shown that sleep benefits the extraction of regularities in the phonological form (e.g. Gomez, Bootzin and Nadel, 2006). In addition, there is evidence suggesting that sleep benefits the extraction of regularities from separately learnt items in the form-meaning *mapping* (e.g. Lau, Alger and Fishbein, 2011), as already discussed in Chapter 1.

These findings that sleep benefits the consolidation of new language mappings can be explained in terms of hippocampal consolidation models, such as the Complementary Learning Systems model (McClelland, McNaughton & O'Reilly 1995; O'Reilly & Norman, 2002), which has been already introduced in Chapter 1. This model suggests that new information initially enters the hippocampus. Then, the new information is gradually integrated with existing knowledge in the neocortex during offline consolidation. Sleep has been shown to play an active role in this offline consolidation, supporting the processes of memory stabilization, strengthening and integration of individual items with existing knowledge, as well as abstraction of regularities (Stickgold & Walker, 2013).

The hippocampal memory system is thought to be initially involved in learning new arbitrary mappings more than non-arbitrary, i.e. systematic mappings (e.g. McClelland et al., 1995). Thus, the aim of the current study was to test the hypothesis that arbitrary form-meaning mappings may benefit from sleep-related consolidation more than systematic mappings, given that arbitrary mappings may be more hippocampally dependent. It was of interest to test this hypothesis, given that previous findings have shown that sleep is likely to benefit the consolidation of hippocampally-dependent information (e.g. Rasch, Buechel, Gais & Born, 2007).

2.2 The Experimental language

The same artificial language was used both in Experiment 1 (i.e. in this Experiment, Chapter 2), and throughout this thesis (i.e. in the language tasks presented in Chapters 3, 4, 5 and in the exploratory correlational analysis in Chapter 6). The artificial language was used to investigate the consolidation of newly learned words and their meanings (e.g. "*tib bisesh"* = "queen"; "*ked cassool"* = "doctor"), specifically whether sleep-related consolidation varies with arbitrariness in the form-meaning mapping. The trained language was modelled on grammatical gender systems, with the more systematic mappings implemented via determiners (e.g. '*tib', 'ked'*) and suffixes (e.g. '*esh', 'ool')* and the natural gender of the referents (see Table 1 for examples). Arbitrary mappings were implemented via individual stem-meaning mappings (e.g. '*bis_'* = "queen", *'cass_'* = "doctor").

In the Experiments presented later in this thesis, we varied the degree of systematicity of the determiner mapping in relation to the suffix mapping. Specifically in Experiment 1 (i.e. in this Chapter), the determiner-gender regularity was a fully systematic mapping as there was one determiner per gender (e.g. '*tib'* = "female"; '*ked'* = "male"). The suffix-gender regularity was a mapping of intermediate systematicity, as there were two suffixes per gender (e.g. '*_esh', '_eem'* = female, '*_ool', '_aff'* = male). Varying levels of systematicity in the language allowed to investigate firstly whether arbitrary mappings (i.e. stems) are processed differently to systematic (in terms of learning, remembering and retrieval processes involved in consolidation). Secondly, it allowed to ask if the sleep state (compared to the equivalent time awake) is particularly beneficial for learning individual specific (arbitrary) items or for extracting regularities (contained within systematic mappings).

Table 1. Example of mappings used.

Gender	Suffix	Determiner	Example Word-Picture Pairs
Feminine	'*_eem*' '*_esh*'	*'tib'*	tib tormeem - bride tib mofeem – waitress tib bisesh - queen tib cherkesh - angel
Masculine	'*_ool*' '*_aff*'	*'ked'*	ked cassool - doctor ked hormool - priest ked darlaff - fisherman ked zarfaff - mechanic

A study using a similar method as in the current experiment (using the same language tasks and items) but with a nap paradigm (Mirkovic & Gaskell, 2016) showed some evidence that following a 90-minute daytime nap the memory of individual vocabulary items was improved unlike the memory for the grammar (determiners, suffixes). Neither the sleep nor the wake group acquired the suffix-gender regularity which was of intermediate systematicity. Although both the sleep and wake groups learnt and could generalize the most systematic determiner-gender regularity mapping, there was no additional benefit of sleep for this. Recall performance was very low for both the wake and the nap groups.

These findings can be explained by the Complementary Learning Systems (CLS) model, to suggest that sleep particularly benefits the consolidation of arbitrary information (vocabulary) compared to systematic information (grammar) (e.g. McClelland et al., 1995).

2.3. Experiment 1

Experiment 1, i.e. the first study in this thesis will investigate the predictions of the CLS model (e.g. McClelland et al., 1995), behaviourally comparing the consolidation following a 12-hour period of *overnight* sleep against consolidation following a 12-hour period of daytime wakefulness. More specifically, we are investigating whether a period of *overnight* sleep will benefit the consolidation of individual arbitrary mappings (i.e. individual stems), but not the extraction of regularities (i.e. the determiner-gender mappings, and suffix-gender mapping regularities).

2.3.1 Method

Participants

Fifty-eight monolingual native English speakers from the University of York participated in the study. Out of these, four were discarded (3 participants due to failure to complete a task and 1 participant did not return for the second session). Participants did not have any language, sleep, psychiatric or neurological disorders, and were not on any psycho-active medication and had no history of drug or alcohol abuse (as measured by questionnaires). Participants received either 2 hours undergraduate psychology course credit or £12 and all gave written informed consent.

Participants refrained from caffeinated products (in drinks or food), smoking, alcohol, drugs and daytime naps in the 24 hours prior to starting the first session and until both sessions were completed. Participants were instructed to sleep for a minimum of 6 hours the night prior to the study, to have woken up before 8 am on the day of the study and to have eaten before the session. Before starting each session, participants completed the Stanford Sleepiness Scale and a reaction time vigilance test to check alertness levels. The sleep group (6 male, 4 left handed) had 27 participants (mean age 19.30 ± 1.27 [SD]). The wake group (5 male, 5 left-handed) also had 27 participants (mean age 20.21 ± 6.11 [SD]). Seven participants from the sleep group and seven from the wake group played a string instrument (this question was included in the questionnaire).

Design

Participants were either in the daytime-wake or the overnight-sleep (at home) groups (between-subjects design). The overnight group was trained in the evening (at approximately 8.30 PM) and tested in the morning (at approximately 8.30 AM). The wake group was trained in the morning (at approximately 8.30 AM) and tested in the evening (at approximately 8.30 PM) as seen in Figure 1. When signing up for the study, participants could choose whether their first session would be in the morning or in the evening (to coincide with their personal schedules).

Time:	8:30 PM	Sleep at home	8:30 AM
Sleep Group	Training		Tests
Time:	8:30 AM	Daytime wake	8:30 PM
Wake Group	Training		Tests

Figure 1. Experiment structure.

Materials

All tasks were presented on a PC using DMDX software except for the 2D object location task which was presented using E-prime 1.0 software. Responses were recorded using the Beyerdynamic headset and a USB buttonbox.

Language tasks' stimuli

All of the language tasks were counterbalanced on the language learnt (in reference to the determiner-natural picture gender mappings) – for half of the participants the determiner '*tib'* and the suffixes '-*eem'* and '-*esh'* mapped onto pictures of human occupations whose natural gender was female and the determiner '*ked*' and the suffixes '-*ool'* and '-*aff'* mapped on to male and vice versa for the other half of the participants (i.e. '*tib'* = male, '*ked'* = female).

There were a total of 16 training and 32 generalisation new artificial words. All of the artificial words used were pronounceable English pseudo-words. They were paired with images of human occupations with corresponding female or male characters selected from the ClipArt database. The words and corresponding pictures together with their sound recordings (with and without the determiners, recorded by a native-English speaker) were selected from Mirkovic, Forrest and Gaskell's (2011) study, also used in Mirkovic and Gaskell's (2016) study (Figure 2). All words had gender determiners (determiners mapped on to the natural gender of their referents, e.g. the determiner '*tib*' mapped on to female pictures, the determiner '*ked'* on to male) and gender suffixes ('*_esh*', '*_eem*' – female; '*_aff*', '*_ool*' - male). The determiners, word stems and suffixes were originally chosen using the ARC nonword database (Rastle, Harrington, & Coltheart, 2002). All words consisted of a 3-5 phoneme, 1 syllable long stem followed by one of the four possible suffixes.

ked cassool **tib bisesh**

Figure 2. Example picture-word pairings used.

Training items

Out of the 16 trained words, 8 mapped onto female pictures, the remaining 8 mapped onto male pictures. In other words, there were 8 different individual stems per gender. As there were 4 suffixes (2 per gender), and 2 determiners (1 per gender), this means that there were 4 stems per suffix, and 8 stems per determiner in the training set.

Generalisation items

For each counterbalanced language, the two generalisation tasks contained a total of 32 new items – 16 per task (32 new stems and 32 new pictures none of which were seen during training). The two tasks (described below) used the same gender determiners and gender suffixes as during training. All of the new pictures also represented human occupations (male and female in gender). Of the new word-picture mappings which participants saw within each of these two tasks, half of the mappings were consistent with the mappings between the determiners and suffixes and the natural gender of the characters that participants were exposed to during training and half were not. These two tasks were named the suffix generalisation task and the determiner and suffix generalisation task the details of which are described below.

The suffix generalisation task

This task tested participants' extraction of the suffix-gender regularities by testing the participants' ability to discriminate between new word-picture pairs which were either consistent or inconsistent with the suffix mappings in the trained language. Therefore, for the inconsistent items the new word had a suffix which was inconsistent both with the determiner and the natural gender of the pictured character. For example, if during training participants learnt that the determiner '*tib'* and the suffix '*esh*' always mapped onto female gender pictures and the suffix '*aff'* onto male gender pictures, then a consistent word-picture mapping would be 'tib sarb*esh*' - 'witch' picture (all female gender), and an inconsistent word-picture mapping would be 'tib lup*aff'*- 'geisha' picture (female gender determiner and picture but male gender suffix). In order to correctly press 'mismatch' to inconsistent word-picture mappings (see more details on the task below), participants could either rely on phonological determiner-suffix or semantic suffix-picture gender mappings extracted during training.

The determiner and suffix generalisation task

This task tested participants' extraction of the mapping between the determiner and the picture gender. Therefore, for the inconsistent items the determiner-gender mapping remained consistent with the trained gender mappings but the (natural) picture gender was inconsistent. For example, an inconsistent word-picture mapping would be '*tib* curr*esh*' - 'policeman' picture (female gender determiner and suffix but male gender picture). In order to correctly press 'mismatch' to inconsistent word-picture mappings, participants could either rely on determiner-picture gender mappings or suffix-picture gender mappings (if participants could correctly distinguish between consistent and inconsistent mappings on this task but not on the suffix generalisation task, then participants relied on the determiner-picture gender mappings in this task, for which reason it was called the determiner and suffix generalisation task).

Procedure

All participants completed tasks over 2 sessions spaced 12 hours apart. The order in which participants completed the tasks is shown in Table 2. As seen in the table, we also included a procedural memory task and a declarative memory task. These two tasks will be included in every experiment in this thesis, but the specific methodology and results of these two tasks are presented in Chapter 6, i.e. separately from the language tasks' results. All vigilance results are presented in the Appendices.

Table 2. Procedure (order of tasks).

Table 2 A: Session1.

Tasks	Stimuli
Vigilance	Vigilance RT task, Vigilance questionnaires (including Stanford Sleepiness Scale)
Declarative memory:	2D Object location exposure and training
Procedural memory:	Finger-tapping exposure
Language training:	a. Repetition task
	b. Word-picture matching
Procedural memory:	Finger-tapping training

Table 2 B: Session 2.

Tasks	Stimuli
Vigilance	Vigilance RT task, Vigilance questionnaires
Declarative memory:	2D object location delayed test
Procedural memory:	a. Finger-tapping retest
	b. Finger-tapping generalisation*
Language tests:	a. Recall
	b. Determiner selection
	c. Suffix & determiner and suffix generalisation**
	d. Translation recognition

*New sequence. **New items. All other tasks contained trained items only.

Information sheets about the experiment were sent out to participants prior to the first session (to inform participants of what was required of them). Before starting the main tasks in both sessions, participants completed the Stanford Sleepiness Scale (SSS) and a vigilance task. At the beginning of session one, all participants were screened for sleep disorders using the Epworth test and the Sleep Habits Questionnaire and completed a sheet about their relevant demographic information (language disorders, foreign language proficiency, country of residence until age 10, musical instrument ability). Over the two sessions, participants completed tasks testing declarative memory (2D object location task, Rasch et al. 2007), procedural memory (finger-tapping, Walker et al. 2002, 2003) and language tasks (whereby the training and generalisation tasks used the word picture matching paradigm, as with Breitenstein, Zwitserlood, de Vries, Feldhues, Knecht & Dobel, 2007).

Session 1 lasted approximately 60 minutes (including consent forms, questionnaires and all tasks); Session 2 lasted approximately 45 minutes (including questionnaires and all tasks). The language tasks used in the experiment are described below. As a reminder, the procedural memory task and the declarative memory task, including the results, are described and presented in Chapter 6. The vigilance RT task and vigilance questionnaires, including the results are also described and presented in the Appendices.

Language training tasks (Session 1)

Repetition

A fixation cross appeared for 500 ms at the centre of the screen. Participants then heard the new word presented through headphones. Participants were asked to repeat each word out loud after hearing it. There were 2 practice trials and 48 main trials (each of the 16 words to be learnt were repeated 3 times in a randomized order). This task was used to familiarise participants with the phonological forms (and therefore was not included in the analysis, as the main aim of the experiment was to test phonological form-meaning *mappings*).

Word picture matching

The training method used probabilistic exposure (as with Breitenstein et al., 2007). This type of initially implicit learning without feedback over several repetitions has been found to be a successful way of acquiring new vocabulary with semantics and retaining it over time for both healthy participants and those with aphasia (Breitenstein & Knecht, 2002).

There were a total of 6 practice and 192 main trials. The main trials were presented over 4 blocks during which participants saw the 16 items (word-picture combinations) a total of 12 times. Overall, the items were paired with the correct picture (of a male or female human occupation) 8 times. These were labelled as target trials in the results section. The other 4 times the items were paired with 4 different incorrect pictures (2 male and 2 female gender pictures). These were labelled as noise trials in the results section. Within each block each of the 16 items was seen twice paired with its correct picture, once with an incorrect (48 word-picture combinations randomly presented). Participants could take a break between each block.

After a 500 ms fixation cross, the new word was heard (presented through headphones, using the same recordings as during the repetition task), and 200 ms later a picture of a human occupation was presented. The picture stayed on screen for a maximum of 1500 ms (i.e. the next trial began as soon as the participant responded or after the maximum time of 1500 ms if there was no response). Participants were asked to indicate using a buttonbox whether each word matched to the picture it was presented with and to answer according to their intuition, "without thinking too much". This task was used to (initially implicitly) train participants on new form-meaning mappings. Participants responded using the buttonbox, and were instructed that the 'right' hand-side button was 'match' and 'left' was 'mismatch'.

Language tests (Session 2)

Most arbitrary mapping: vocabulary

Translation recognition task

This task was used to assess participants' recognition memory of individual trained vocabulary items. There were 4 practice and 32 main trials. After a 500ms fixation cross, an English word was presented over the headphones. This was followed after an 800 ms pause by one of the trained artificial words (without the determiners). Participants were asked to respond if the two words (the English word and the artificial word) 'matched' or 'mismatched' using the buttonbox (pressing 'right' for 'match' and 'left' for 'mismatch'). Of the 32 main trials, 16 trials 'matched' and 16 did not. The 'mismatching' items were designed such that participants could not use their knowledge of suffix-picture gender regularities if it was extracted (for example 'scoiff*eem*'-'nurse' pairing was a 'match' as it was trained but 'scoiff*eem*'-'queen' pairing was a 'mismatch' despite the correct suffix-picture gender mapping).

Recall task (picture naming)

This task was used to assess participants' knowledge of trained suffixes and individual vocabulary items. There were 2 practice trials (i.e. one picture presented twice) and 32 main trials (i.e. 16 different words). Each picture was presented twice in a row. The first time participants saw the picture only. The second time participants saw the picture and the first grapheme written underneath it, representing the first *sound* of the word to be recalled (e.g. if the word was 'bisesh' the first grapheme was 'b' and if the word was ‘phlaveem’ the first grapheme was ‘ph’). These were used as examples to encourage participants to name the pictures without the determiners.

After a 500ms fixation cross, the picture remained on screen for 5,000 ms. Participants were asked to name the picture out loud (by trying to recall one of the 16 words learnt during training) while it remained on screen. Participants’ responses were recorded (using a headset), and the transcriptions of these recordings were used for analysis.

Systematic mappings: grammar

Suffix generalisation, and determiner and suffix generalisation tasks

The two tasks were used to assess participants’ extraction of suffix-gender regularities (suffix-determiner mapping and suffix-picture gender mapping) and determiner-picture gender mapping regularities. There were 6 practice and 16 main trials in each of the 2 tasks. The same word-picture matching task was used for these two tasks as with training, where participants were asked to decide whether the sound/word matched with the picture or not by pressing the buttonbox. The order of the presentation of the two generalization tasks was counterbalanced across participants.

Most systematic mapping: Determiner Selection

This task was used to assess participants’ recognition memory of the individual trained determiners. There were 4 practice and 16 main trials. After a fixation cross, both of the determiners ‘KED’ and ‘TIB’ appeared on either side of the screen and remained for a maximum of 1,500 ms (or until a response if it occurred before the 1,500 ms). Simultaneously, (during the first 500 ms of each trial) one of the trained words was heard (using the same recordings as for the translation task, i.e. without determiners) and a picture appeared on screen until the response was provided, together with the determiners (to be selected from) ‘KED’ and ‘TIB’. Participants were asked to indicate if each word was a ‘*ked*’ or a ‘*tib*’ word by using the buttonbox.

The determiner location on the screen was counterbalanced (i.e. for half of the participants the determiner ‘*tib*’ appeared on the left hand-side of the screen as the determiner ‘*ked’* appeared on the right, for the other half ‘*tib*’ was on the right while ‘*ked’* was on the left, participants used the corresponding buttons (‘left’ and ‘right’) on the buttonbox to respond).

2.3.2 Results

Language tasks

Training task (Session 1)

Participants were trained on the artificial language using a word-picture matching task. In the training task, each participant learnt one of two language versions which were counterbalanced for the mapping between the determiners and suffixes and the natural gender of the depicted character. Both the accuracy (described first) and the RT analysis (described in the section below, and included to illustrate the speeding up of RTs across learning) for the training task will be presented.

Accuracy analysis

Participants' accuracy at the end of training (final block of the task) was measured to assess their level of learning of the language. It was expected that by the end of training, there would be no significant differences between the groups (sleep or wake) or language versions, given that this was the participants' first exposure to the novel language. A two way between subjects ANOVA was run with the independent variables of group (sleep or wake) and language (version one or two). The dependent variable was accuracy (proportion of correct responses) at the final block of the training task. There was no significant effect of group ($F(1, 50) = 0.52$, $p = .477$) or of language version ($F(1, 50) = 0.48$, $p = .492$) and no significant interaction between group and language ($F(1, 50) = .95$, $p = .336$). As shown in Figure 3, participants from both groups (sleep and wake) and both language versions reached a similar level of performance by the end of training.

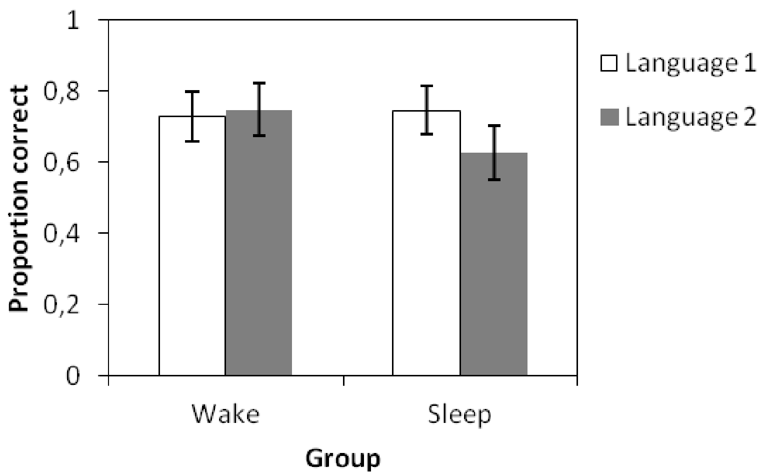


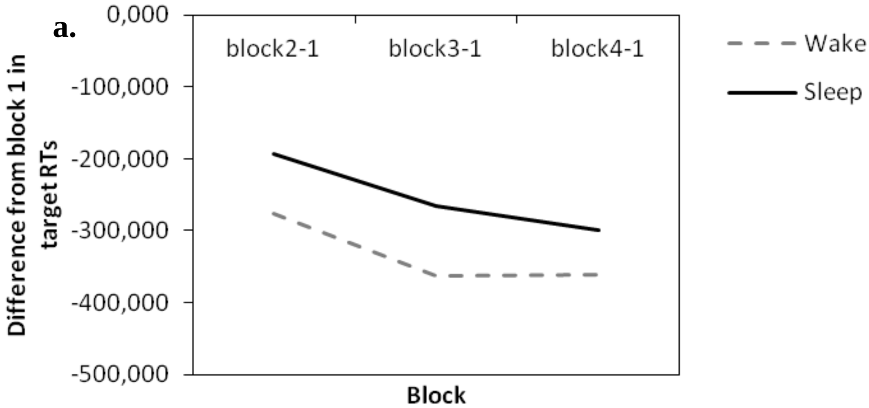
Figure 3. Performance on the last block of the training task for the two groups and the two versions of the training language. NB. In this Chapter and throughout this thesis, the standard error bars in the graphs represent the standard error of the mean.

Correct RTs analysis on the training task

The RT data for correct responses was analysed separately for target trials and separately for noise trials, because of the different number of data points contributing to the analyses. As described in the method (the word picture matching task, procedure), the number of trials for the target pairings was greater than for the noise trials. The reaction time analyses will show the expected speeding up with increased exposure. By the end of the training task, participants should acquire the target mappings. As a reminder, each word was seen paired with its 'target' picture a total of 8 times (e.g. the pairing "tib bisesh – queen" was seen paired together 8 times). These 8 trials were labelled the target trials. Each word was seen another 4 times paired with 4 different 'noise' pictures, 2 male and 2 female gender pictures (e.g. "tib bisesh – angel" (female), "tib bisesh – waitress" (female), "tib bisesh –builder" (male), "tib bisesh – priest" (male)). These 4 trials were labelled the noise trials.

A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake and sleep). First, the dependent variable was the change in target RTs compared to block 1. There was a significant main effect of block ($F(1, 49) = 20.04$, $p < .001$), no significant main effect of group ($F(1, 49) = .21$, $p = .650$), and no significant interaction between block and group ($F(1, 49) = 1.40$, $p = .252$), as seen in Figure 4.

a.



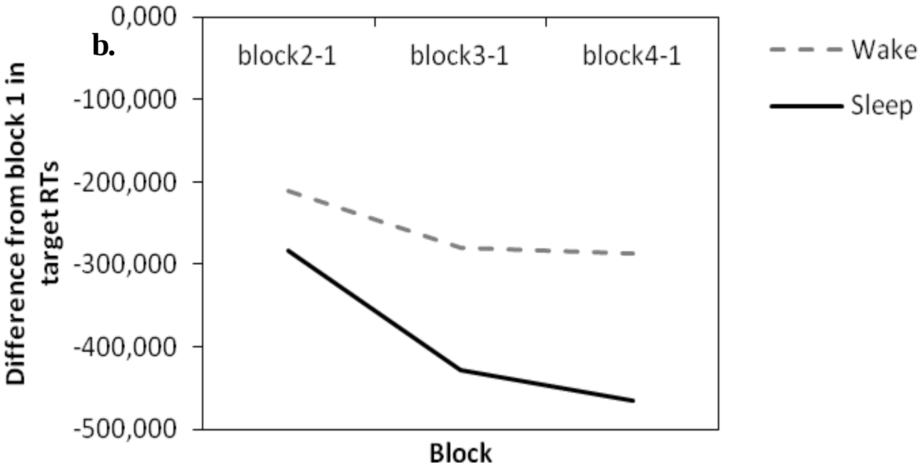


Figure 4. Difference in target RTs from block 1 (Figure 4a. shows language version 1 RTs, Figure 4b. shows language version 2 RTs).

There was also no significant main effect of language version (F(1, 49) = .33, p = .569) and a marginally significant interaction between group and language (F(1, 49) = 3.57, p = .065). There was no significant interaction between block and language (F(1, 49) = .46, p = .633), and no significant interaction between block, group and language (F(1, 49) = .88, p = .419). As shown in Figure 4, all participants improved across the blocks, regardless of group. In addition, as shown in Figure 4, and as reflected by the marginally significant group x language interaction, participants in the wake group improved more in language version 1 (Figure 4a), but participants in the sleep group improved more in language version 2 (Figure 4b).

Secondly, the dependent variable was the change in noise RTs compared to block 1. One participant was not included in the analysis, as they did not have any correct responses for noise RTs in some of the blocks. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake and sleep). The dependent variable was the change in noise RTs compared to block 1. There was a significant main effect of block (F(1, 48) = 10.19, p < .001), no significant main effect of group (F(1, 48) = .48, p = .494), and no significant interaction between block and group (F(1, 48) = 1.51, p = .227), as seen in Figure 5.

There was also no significant main effect of language version (F(1, 48) = 1.08, p = .303) and no significant interaction between group and language (F(1, 48) = 1.49, p = .228). There was no significant interaction between block and language (F(1, 48) = .28, p = .755), and no significant interaction between block, group and language (F(1, 48) = .97,

p = .383). All participants showed similar improvement across the blocks, regardless of group (as seen in Figure 5), and regardless of language version.

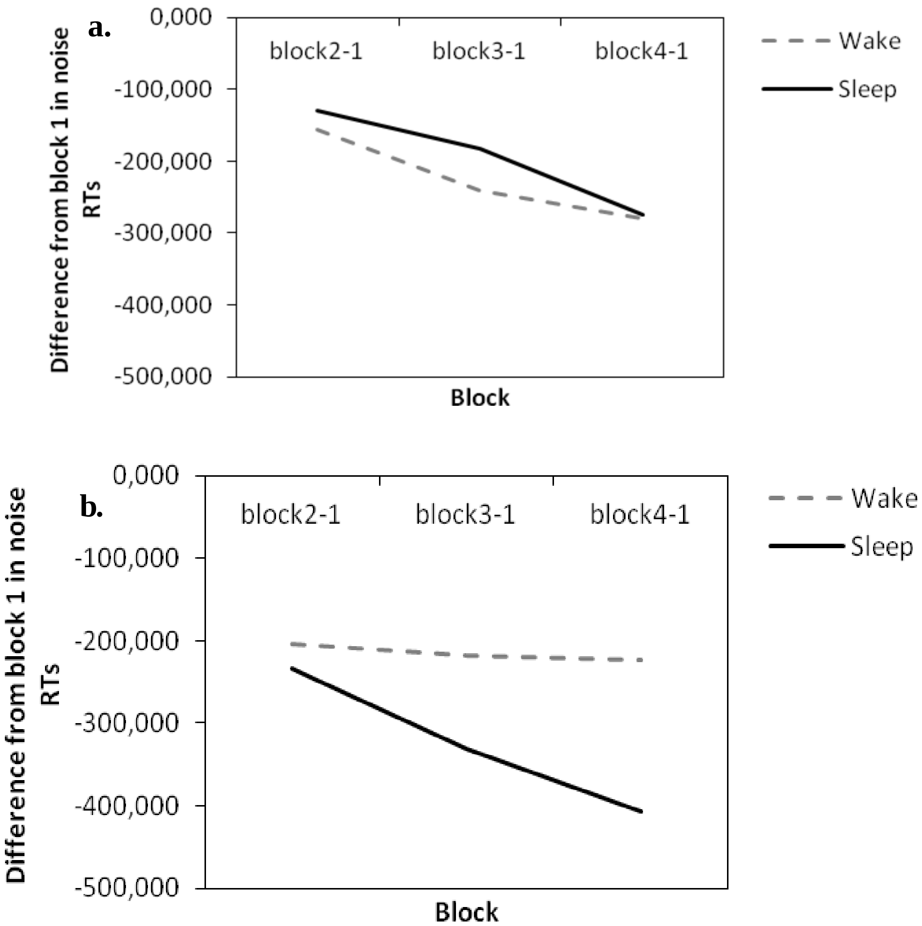


Figure 5. Difference in noise RTs from block 1 (Figure 5a. shows language version 1 RTs, Figure 5b. shows language version 2 RTs).

Overall, the findings from the training task suggest that all participants reached a similar level of performance in terms of accuracy by the end of training, regardless of group. This means that any differences found between the groups after a consolidation delay including sleep or wake in Session 2 cannot be due to differences in initial acquisition levels. Given that there were no significant differences between the two versions of the language all further analyses will be collapsed across the versions. The RT analyses suggested that all participants speeded up across the exposures, as expected, indicative of learning of the language. Given that there were no main effects or interactions with language version in the noise RTs and accuracy analyses, the analyses below collapse across this variable (i.e. the variable of language version), and focus on the differences between groups.

Language tests (Session 2)

Most arbitrary mapping: vocabulary

Two different tests were used to measure vocabulary, i.e. the arbitrary mappings. Specifically, the translation recognition task (presented first), and the recall task.

Translation recognition task

Participants' memory of trained vocabulary was tested using a translation recognition task, which consisted of the auditory presentation of the novel words together with (matching or mismatching) English translation equivalents for the novel words (e.g. 'queen' for '*bisesh*'). Recognition memory was assessed using the non-parametric discrimination (A') formula (Donaldson, 1992). This formula took into account firstly participants' discrimination ability (between same i.e. matching and different i.e. mismatching item pairs) and secondly participants' response bias (to press 'yes' or 'no').

For example, participants learnt during training that '*bisesh'* meant 'queen'. Correct responses would be pressing 'yes' for the 'queen - *bisesh*' (match) trial and 'no' for the 'queen - *scoiffesh*' (mismatch) trial. Incorrect responses would be pressing 'no' for 'queen - *bisesh*' (match) and 'yes' for 'queen -*scoiffesh*' (mismatch). Chance level discrimination would be indicated by an A' of 0.5 and perfect discrimination would be indicated by an A' close to 1.0.

A between subjects t-test was run with the independent variable of group (sleep or wake) and A' discrimination as the dependent variable. There were no significant differences between groups ($t(52)= -1.32$, $p = .192$). As shown in Figure 6, participants had an overall good level of discrimination between match and mismatch pairs. In addition, numerically the accuracy averages reflected that participants in both groups reached a good level of performance. For the wake group, the mean accuracy was 69.9% for match trials, and 82.4% for mismatch trials. For the sleep group, the mean accuracy was 78.0% for match trials, and 85.6% for mismatch trials.

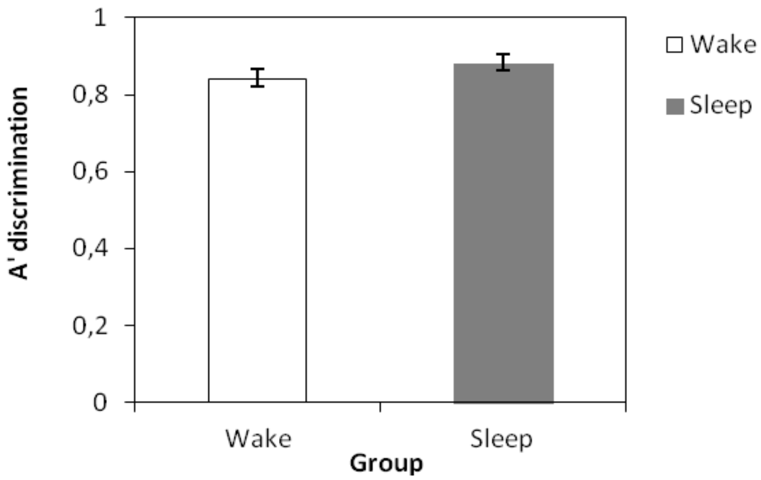


Figure 6. Performance on the translation recognition task.

Recall (stems)

Participants' memory for trained vocabulary was also measured using a picture naming recall task. The task required the participant to recall the novel words at a prompt which contained either the picture with which the word had been trained, or a picture together with the initial grapheme (e.g. 'b' for '*bisesh*'). As words were made up of stems and suffixes (e.g. stem: *bis_*, suffix: *_esh*), to measure the memory of the mappings at the most arbitrary level, from the words participants produced just the stems were extracted for the analysis of stem accuracy, regardless of the suffix accuracy. For example, for the trained '*bisesh*- queen' mapping, on seeing the picture 'queen', participants' production of e.g. 'bis*eem*' was still marked as correct for stem accuracy analysis as 'bis' was the trained stem.

If sleep benefits the memory of the arbitrary mappings as expected, participants in the sleep group should recall more stems correctly than participants in the wake group. A mixed ANOVA was run with the between-subjects variable of group (sleep or wake) and the within subjects variable of cue type (picture only or picture and grapheme). The dependent variable was stem accuracy (proportion correct out of total productions). Greenhouse-Geisser corrected values are reported because Levene's test was significant for cue type ($F = 4.33$, $p =. 042$), indicating unequal variance.

Both groups of participants recalled a very small proportion of the novel word stems (Figure 7). However, participants in the sleep group performed significantly better than participants in the wake group as evidenced by a significant main effect of group ($F(1, 52) = 18.60$, $p <.001$). There was no significant main effect of cue type ($F(1, 52) = 1.53$, $p = .222$), and no significant interaction between group and cue type ($F(1, 52) =$

1.18, p = .283). As Figure 7 shows, participants in the sleep group produced more stems correctly than participants in the wake group, as expected, regardless of the cue type.

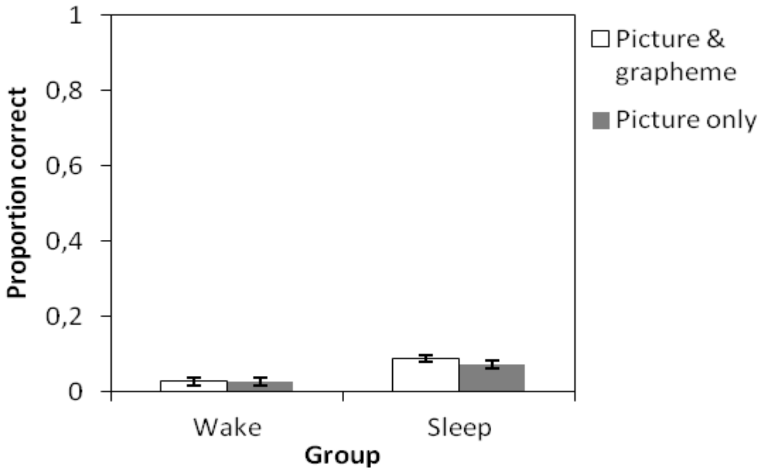


Figure 7. Stem accuracy on the recall task.

The findings from the vocabulary tests suggest that a 12-hour delay between training and testing which included a period of sleep was beneficial for the memory of individual arbitrary sound-meaning mappings compared to the same period awake when measured by a picture naming recall task. There was no evidence of a sleep benefit on the recognition task.

Intermediate systematicity mapping: suffixes (grammar)

Recall (suffixes)

Performance on the recall task was also used to compare the two groups at the intermediate systematicity level (suffixes). It was of interest whether there would be evidence of a sleep benefit for suffix recall. From the words participants produced, just the suffixes were extracted for analysis (i.e. regardless of the stem accuracy – for the trained '*bisesh*-queen' mapping, on seeing the picture 'queen', participants' production of e.g. '*scoiffesh'* was correct for suffix analysis only).

Overall participants in both groups recalled more suffixes correctly than the stems (Figure 8). A mixed ANOVA with the between subjects factor of group and the within subjects factor of cue type was run with suffix accuracy (proportion correct) as the dependent variable. There was a significant main effect of group (F(1, 52) = 27.17, p < .001) and of cue type (F(1, 52) = 15.43, p <.001). There was no significant interaction between group and cue type (F(1, 52) = 1.43, p = .238). As Figure 8 shows, participants in both groups produced more suffixes correctly when the words were cued using a picture and a grapheme. More importantly, participants in the sleep group produced more suffixes correctly than participants in the wake group.

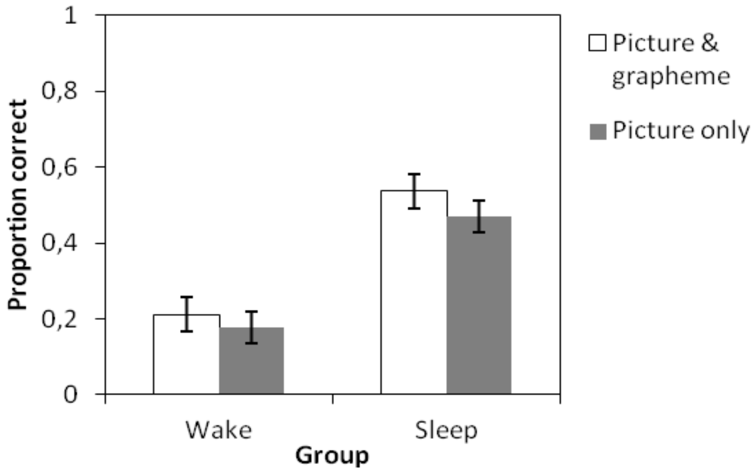


Figure 8. Suffix accuracy on the recall task.

These findings suggest that a 12-hour delay between training and testing which included a period of sleep was beneficial for the memory of individual sound-meaning mappings of intermediate systematicity compared to the same period awake.

Analysis of errors on the suffix recall task

Next, suffix errors were analysed to investigate whether the sleep related memory benefits were likely to be due to individual item memory or rule memory. Rule memory would be reflected by suffix errors indicating that participants predominantly encoded correct suffix-gender mappings (e.g. '*esh*'/'*eem*' but not '*ool*'/ '*aff*' with female picture gender). Participants learnt in total four suffixes during training. In the recall task, when producing a word, a suffix could be incorrect for the specific item but could still be one of the other three suffixes learnt during training.

These suffix errors (productions of trained suffixes which were incorrect for the specific word) were analysed to determine suffix-gender mapping accuracy. A correct suffix-gender mapping would be producing a feminine suffix for a picture of a female human occupation. An incorrect suffix-gender mapping would be producing a masculine suffix for a picture of a female human occupation. For example, during training pictures of female characters were presented with words with the suffixes '*esh*' and '*eem*'. If on seeing the picture 'queen' (correct novel word: '*bisesh*') participants who produced an incorrect suffix (i.e. not '*esh*') were still more likely to produce a feminine suffix '*eem*' than the masculine suffixes '*ool*' or '*aff*' this would indicate that even if participants did not encode the individual sound-picture mapping, they did encode the correct suffix-gender mapping.

Suffix errors reflecting correct suffix-gender mappings (proportion of correct gender suffixes out of the incorrect but trained suffixes) for both picture only and picture and grapheme cue types were compared against the chance level of 0.33, as each produced word could have three possible incorrect trained suffixes and only one of the three could be of the correct gender. Abstraction of the suffix to picture gender mapping would be reflected by correct suffix-gender mappings being produced above chance.

In the wake group, fourteen participants in the picture and grapheme cue type and sixteen participants in the picture only cue type were excluded and in the sleep group four participants from each cue type were excluded from this analysis, because none of their incorrect responses included one of the trained suffixes. Thus in the wake group, the analysis was done on thirteen participants in the picture and grapheme cue type and eleven participants in the picture only cue type. In the sleep group, the analysis was done on twenty-three participants in each of the cue types.

One sample t-tests against chance revealed that for participants in the sleep group the proportion of suffix errors reflecting correct suffix-gender mappings was significantly higher than chance for the picture only cue type, but not for the picture and grapheme cue type. For participants in the wake group, errors were at chance regardless of cue type (see Table 3 for results of the one sample t-tests). These findings provide some evidence for abstraction of the grammatical rule at the intermediate level of systematicity for the participants in the sleep group as errors for correct suffix-gender mappings were produced above chance for the picture only cue type, suggesting a suffix-picture gender mapping was abstracted. Chance performance for the picture and grapheme cue type may suggest that the picture activated gender rule memory whereas the grapheme activated individual item memory.

Table 3. Results of correct suffix-gender errors analysis in the recall task using one sample t-tests.

		t	df value	*p*-value	Mean proportion
Sleep group	Picture and grapheme (N=23)	.47	22	.645	.36

	Picture only (N=23)	2.71	22	.013	.53
Wake group	Picture and grapheme (N=13)	-1.06	12	.309	.24
	Picture only (N=11)	.172	10	.867	.35

Suffix generalisation task

To test participants' ability to apply the suffix-gender regularities they learnt during training to unseen items the suffix generalisation task was used. This was a word-picture matching task with the same methodology as the language training task in Session 1 but using unseen items (unseen word stems and pictures, but with trained determiners and suffixes and the same human occupation genders). Half of the items in this test were consistent with the suffix mapping regularities participants were exposed to during training (e.g. feminine determiner, feminine suffix - female human occupation picture, e.g. *'tib sarbesh'* - 'witch'), half were inconsistent (e.g. feminine determiner but masculine suffix - female human occupation picture, e.g. '*tib lupaff'*- 'geisha'). Participants' endorsement (i.e. pressing 'match') to both consistent and inconsistent word-picture pairs was measured. Crucially, in this task we are interested in the pattern of endorsement of consistent versus inconsistent items, as indication of generalisation ability. Consequently, the most informative results will be the main effect of consistency, and the consistency x group interaction (to indicate whether generalisation varies as a result of sleep versus wake-dependent consolidation).

A mixed ANOVA was run with the between subjects variable of group (sleep or wake) and within subjects variable of consistency (items consistent or inconsistent with trained regularities). Proportion endorsed (pressing 'match' to consistent and inconsistent items) was the dependent variable. There no significant main effect of consistency (F(1, 52) = 1.74, p = .193), a significant main effect of group (F(1, 52) = 4.04, p = .050), and no significant interaction between group and consistency (F(1, 52) = . 89, p = .351), hence there was no indication of suffix-gender regularity extraction.

Next, to investigate this further, the endorsement of consistent and inconsistent items was analysed against chance. If the suffix-gender regularity were extracted the consistent items should be endorsed significantly *above* chance. The inconsistent items should be endorsed significantly *below* chance (i.e. less than .50), if the mapping was fully extracted, or at least *at* chance indicating some extraction.

One sample t-tests against chance (0.5) revealed that participants in the sleep group endorsed both consistent and inconsistent items above chance (consistent: $t(26) = 4.86$, $p = <.001$, inconsistent: $t(26) = 4.54$, $p <.001$). This pattern in the sleep group was not in line with that expected had there been extraction. This is not in line with the findings from the suffix recall error analysis, where there was some indication that the sleep group may have extracted some regularity. Participants in the wake group only endorsed consistent items above chance, and were at chance for inconsistent items (consistent: $t(26) = 2.98$, $p = .006$), inconsistent: $t(26) = 1.66$, $p = .111$). This pattern in the wake group is in line with suffix-gender regularity extraction.

As Figure 9 shows, participants in the sleep group were more likely than those in the wake group to endorse items *regardless* of consistency. The sleep group participants' tendency to endorse more of the previously unseen items than the wake group participants' may reflect that participants in the sleep group were able to remember the correct phonology of the determiners and suffixes, but not the mappings of determiners and suffixes on to picture gender. While there was no *strong* evidence of suffix-gender regularity extraction, the wake group's pattern of endorsement was more in line with suffix-gender regularity extraction (on the basis of the one sample t-tests against chance). Having said this, the wake group did not show evidence of suffix-gender regularity extraction in the error analysis in the recall task.

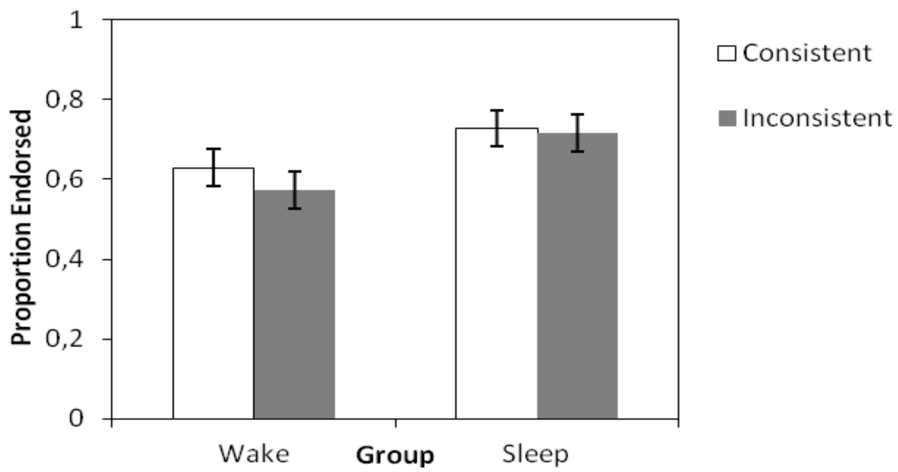


Figure 9. Performance on the suffix generalisation task.

To summarize, the findings from the suffix tests suggest that a 12-hour delay between training and testing which included a period of sleep was beneficial for both consolidating the individual suffixes and abstracting suffix-picture gender regularities. However, there was no evidence to suggest that this was enough to result in the ability to apply the consolidated suffix-gender regularities to unseen items. While there was no *strong* evidence of suffix-gender regularity extraction, it could be *tentatively* suggested that the performance of the wake group on the suffix generalisation task was more in line with suffix-gender regularity extraction, than the performance of the sleep group.

However, there was no evidence of the wake group extracting the suffix-gender regularities in the error analysis in the recall task. In addition, there was no consistency x group interaction in the generalization task to indicate that the wake group's performance was more in line with extraction than the sleep group's performance. Consequently, there was no clear evidence of either wake or sleep benefit for the extraction of suffix-gender regularities.

Most systematic mapping: determiners (grammar).

Determiner selection task

Participants' memory for grammatical regularities which included the determiners (determiner-picture and determiner–suffix mappings) for the trained items was assessed using a determiner selection task. Participants' accuracy in selecting the correct determiner that matched to a picture was assessed. A between subjects t-test was run with the independent variable of group (sleep or wake) and the dependent variable of accuracy (proportion correct). There were no significant differences between groups ($t(52) = -1.67$, $p = .101$). As Figure 10 shows, participants in both groups reached a similarly good level of performance on this task. This indicates that participants in both groups either learned the individual determiner-picture mappings well, or the determiner0gender regularities well. The determiner and suffix generalisation task will clarify this result. There was no evidence for an additional benefit of sleep for determiner consolidation.

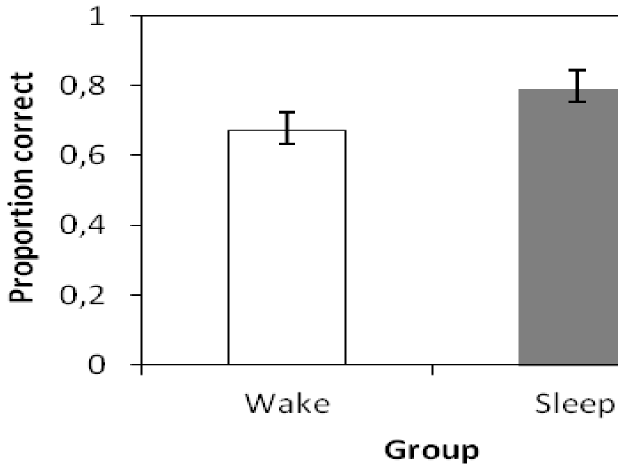


Figure 10. Participants' memory for trained determiners assessed using the determiner selection task.

Determiner and suffix generalisation task

Participants' ability to apply the determiner –picture gender mapping regularities to which they were exposed during training to unseen items was tested using a determiner and suffix generalisation task (another word picture matching task with previously unseen words and pictures). Half of the new items were consistent with the gender mapping regularities from the training set, half were inconsistent. For the inconsistent items, the determiner and the suffix matched each other on trained gender regularities but not to the picture gender (e.g. feminine determiner and feminine suffix with a male human occupation picture, e.g. *'tib curresh'* - 'policeman'). Participants' endorsement (i.e. 'match' responses) to consistent and inconsistent items was measured. As with the suffix generalisation task, the pattern of endorsement of consistent versus inconsistent items is of interest here, as indication of generalisation ability.

A mixed ANOVA was run with the between-subjects variable group (sleep or wake) and the within-subjects variable consistency (items consistent versus inconsistent with the trained regularities). The dependent variable was the proportion endorsed. There was a significant main effect of consistency ($F(1, 52) = 21.66$, $p < .001$), no significant main effect of group ($F(1, 52) = 1.09$, $p = .301$), and no significant interaction between group and consistency ($F(1, 52) = .66$, $p = .421$).

As shown in Figure 11, participants' (in both the sleep and wake groups) stronger endorsement of consistent items over inconsistent items suggests that participants in both groups were able to apply the learnt regularities to unseen items. There was no evidence for an additional benefit of sleep.

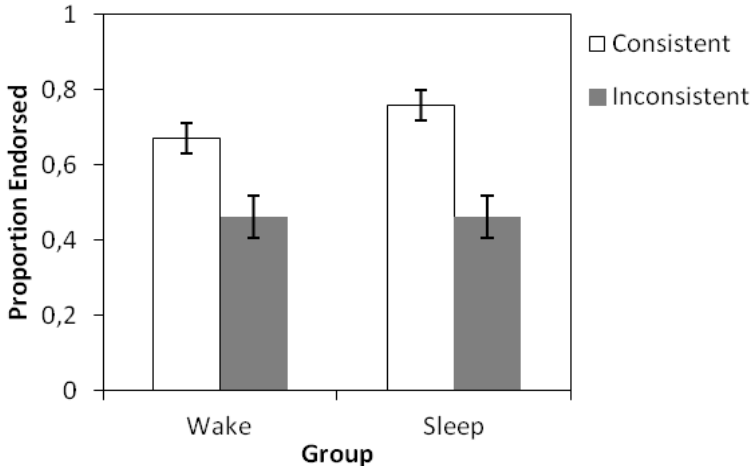


Figure 11. Participants' ability to generalise picture gender mappings assessed using the determiner and suffix generalisation task.

One sample t-tests against chance (0.5) revealed that participants in both groups were significantly more likely than chance to endorse the items consistent with the trained regularities and not significantly more likely than chance to endorse inconsistent items (see Table 4 for results). This pattern is in line with determiner-gender regularity extraction in both groups.

Table 4. Results of one sample t-tests against chance on the determiner and suffix generalisation task.

	Word-picture gender pairing	t	df value	*p* - value
Sleep group	Consistent	6.48	26	<.001
	Inconsistent	-.59	26	.563
Wake group	Consistent	4.10	26	<.001
	Inconsistent	-.81	26	.425

The findings from the determiner tests suggest that participants in both groups learnt the individual determiner-picture mappings and the determiner-gender regularities well and were able to apply the regularities new items. There was no evidence for an additional benefit of sleep for consolidation of these regularities.

2.4. Chapter Summary and Discussion

The first study in the thesis examined whether any differences in the consolidation of new form-meaning mappings emerged following wake-dependent compared to sleep-dependent consolidation. Specifically, we investigated the consolidation of individual, arbitrary mappings and the extraction of regularities contained within systematic mappings. It was found that at the level of the vocabulary (i.e. the arbitrary mapping) the sleep benefit was only seen in the recall task, but not in the recognition task. Findings about the consolidation of the mappings at the intermediate level of systematicity (i.e. the suffixes) were mixed. The suffix recall task provided evidence that sleep was beneficial for the consolidation of suffixes, and also that only the sleep group, but not the wake group was able to abstract the suffix-gender regularity. However, this abstraction in the sleep group was not applied to the suffix generalisation task.

Overall, there was no *strong* evidence of suffix-gender regularity extraction in *either* of the groups. However, it is *tentatively* suggested that the performance of the wake group was more line with suffix-gender regularity extraction. However, this was not confirmed by the analysis of errors in the suffix recall task. Specifically, there was no evidence of the wake group having abstracted the suffix-gender regularity on the analysis of errors in the suffix recall task. In addition, there was no significant consistency x group interaction in the suffix generalization task to indicate that the wake group's performance was more in line with suffix-gender regularity extraction than the sleep group's performance. The fully systematic determiner-gender regularity mapping was extracted regardless of sleep, with no additional benefit of sleep.

Overall, the findings provided some evidence that sleep supports the consolidation of arbitrary mappings. For the intermediate levels of systematicity (i.e. the suffixes), the evidence (from the current study) was mixed. There was no evidence that sleep supports the consolidation of fully systematic mappings (i.e. determiners).

Individual form-meaning mappings

The findings from the recall task may suggest that individual item information (individual phonological forms: stems and suffixes) was used alongside relational information (memory for form-meaning mappings) to retrieve the recalled words. This is consistent with the view that the episodic memory traces which are accessed during recall contain both individual item-specific and relational information (e.g. Smith & Hunt, 2000). The relatively higher recognition accuracy may possibly explain why sleep benefits were not observed on the recognition task, as previous studies with adults have also shown high recognition accuracy even immediately after training (e.g. Dumay & Gaskell, 2007; Gaskell & Dumay, 2003). It may also suggest that recognition memory may rely less on sleep-dependent consolidation, and wake consolidation may be sufficient, consistent with other studies which also have not find any evidence that sleep is beneficial for recognition memory (e.g. Brown et al., 2012).

Individual form-meaning mappings and the CLS

We found that at the level of individual new words, as predicted, sleep benefited the consolidation of the most arbitrary mapping at the level of individual items (stems) and of suffixes and their meanings. This suggests that sleep benefits the consolidation of individual arbitrary form-meaning mappings, consistent with hippocampal consolidation models, such as the CLS model (McClelland et al., 1995). As previously described, this model would suggest that the individual form-meaning mappings were initially more reliant on the fast-mapping hippocampus than the slow-mapping neocortex during encoding. After a period of sleep-related consolidation, these hippocampus-dependent mappings would be stabilized for more permanent storage, becoming less reliant on the hippocampus for retrieval.

Consistent with evidence suggesting that sleep facilitates the process of offline consolidation of newly learned words (e.g. Brown, Weighall, Henderson & Gaskell, 2012; Dumay & Gaskell, 2007; Henderson, Weighall, Brown & Gaskell, 2012; Tamminen & Gaskell, 2008; Williams & Horst, 2014) we found that sleep provided a benefit (over the equivalent time awake) for the consolidation of individual arbitrary form-meaning mappings.

Regularities in the form-meaning mappings

Experiment 1 provided no clear evidence of a sleep benefit for the extraction of regularities, unlike previous findings with infants (Gomez et al., 2006) who only learned phonological forms. This discrepancy may be due to age-related differences and processes. Children have greater amounts of slow wave sleep (SWS) than adults (Anders et al., 1995; Ohayon et al., 2004; Marshall & Born, 2007) – a stage of sleep which would be expected to be involved in consolidation of new vocabulary and new grammar (Born et al., 2006; Walker & Stickgold, 2004). Children also have different neural memory systems to adults, showing increased plasticity of neural structures (e.g. Menon et al., 2005).

The findings of Experiment 1 also do not support the previous finding in *adults* for a benefit of sleep for the learning of new form-meaning mappings containing overlapping regularities (Lau et al., 2011). However, in Lau et al.'s (2011) study both the participants who slept, *and* those who stayed awake performed similarly on firstly the cued recall task and secondly the recognition of trained items (with the delayed-wake group even having better recognition performance of English meanings for the trained characters than the delayed-sleep group). This suggests that (unlike in our Experiment 1) in Lau et al.'s (2011) study sleep did not boost consolidation of *individual* items. Consequently, sleep may have preferentially benefited the consolidation of regularities but *not* the individual items in Lau et al.'s (2011) study, which is the opposite to the findings from our Experiment 1.

This may be important, given that other previous findings have indicated that the stabilization of individual elements *may* inhibit the generalisation of trained regularities, and consequently some forgetting of the individual elements may be needed for extraction and hence generalisation to emerge (e.g. Vlach, Ankowski, & Sandhofer, 2012; Werchan & Gomez, 2014).

Other key differences in comparison to our Experiment 1, are that participants in Lau et al.'s (2011) study learnt Chinese characters which belonged to *seven* semantic groups, whereas in our Experiment 1 participants were exposed to only *two* semantic categories (male and female). This may suggest that increased variability (for example of semantic categories) may boost sleep-dependent benefits for the extraction of regularities, whereas decreased variability may boost sleep-dependent benefits for consolidation of individual specific items. This is also confirmed by findings from Gomez et al.'s (2006) study, as the nap-control group who learned items with less variability (3 different middle items compared to 24) compared to the experimental nap group also did not show evidence that sleep boosts the extraction of regularities.

Sleep parameters underlying the behavioural sleep benefit

In Experiment 1 (i.e. in this Experiment), there was evidence for a benefit of sleep for the consolidation of arbitrary mappings after just one night. Therefore, it may be that the consolidation of arbitrary versus systematic mappings is also mediated by different sleep stages and parameters. Indeed, previous evidence suggests that slow wave sleep (SWS) may be beneficial for replay of previously acquired memories such that they can be strengthened (e.g. Born, 2010). Moreover, there is evidence to suggest that SWS benefits more hippocampally-dependent information (e.g. Rasch et al., 2007) and on the basis of the CLS (e.g. McClelland et al., 1995) we are suggesting that our individual arbitrary mappings may be more hippocampally-dependent, and consequently their consolidation may be facilitated by SWS.

Therefore, Experiment 2 (i.e. the next study) will use polysomnography in order to ask more detailed questions about the nature of sleep which mediates the benefit found for the consolidation of arbitrary information. In particular, it will be of interest to correlate SWS with the recall and generalisation tasks in order to determine the involvement of SWS for arbitrary versus systematic information.

Conclusions

In conclusion, in Experiment 1 we found evidence that sleep benefitted the consolidation of *individual* form-meaning mappings at the most arbitrary mapping (vocabulary), but not the most systematic mapping (determiners), which was learned well without sleep. The evidence for suffix-gender regularity extraction was mixed, possibly because the suffix-gender regularity was a less systematic mapping than the determiner-gender regularity. The difference between determiners and suffixes and the factors that contribute to the extraction of regularities will be explored in later Chapters.

The next Chapter will ask whether SWS is associated with the benefit of sleep that was found for the consolidation of the individual form-meaning mappings, with the expectation that SWS will benefit individual arbitrary mapping consolidation.

Chapter 3: Slow wave sleep (SWS) and the consolidation of arbitrary mappings

3.1 Introduction

In Experiment 1 we found a *behavioural* benefit of sleep for the consolidation of arbitrary mappings. There was no strong evidence for a benefit of sleep in the extraction of regularities (determiners-gender and suffix-gender regularities, as measured by two generalisation tasks). Moreover, the evidence for suffix-gender regularity extraction in Chapter 2 (Experiment 1) was mixed. Experiment 2 aims to extend these behavioural findings, by investigating whether SWS underlies the behavioural benefit of sleep for the consolidation of arbitrary mappings found in Experiment 1. To achieve this goal, Experiment 1 will be partially replicated (no wake group will be included) and both behavioural and polysomnographic data will be collected.

As already discussed in Chapter 2, we have based our predictions on the basis of the CLS model (e.g. McClelland et al., 1995). When applied to language learning and consolidation, the CLS would predict that sleep benefit the consolidation of arbitrary, but not systematic mappings. This is indeed what we found in Experiment 1. According to the CLS, this is because arbitrary mappings are more hippocampally-dependent than systematic mappings. Indeed, there is evidence to support this. For instance, (as already described in Chapter 1) Bretenstein et al. (2005) have found that the consolidation of individual, arbitrary word-picture mappings (such as "bini-book") is related to the activation of the hippocampal system. This has also been supported by other studies (e.g. Takashima et al., 2006 and Takashima et al., 2009).

There is in addition some evidence to suggest that sleep, and SWS more specifically, may mediate the hippocampal-neocortical dialogue necessary for the consolidation of the arbitrary mappings (e.g. Marshall & Born, 2007). To elaborate on SWS as a sleep parameter, the four different stages of sleep are differentiated on the basis of differential brain and muscle activity. Stages 3 and 4 in combination are termed SWS, which is the deepest stage of sleep. SWS is typically associated with high amplitude delta wave activity between 1 and 4 HZ, and slow oscillations at a frequency of approximately 0.75 HZ, as well as sleep spindles (Steriade, 2003).

There is evidence to suggest that the different stages of sleep are typically associated with different behaviours and memories, as discussed in Chapter 1. More specifically, SWS is typically associated with declarative information consolidation (e.g. Diekelmann & Born, 2010; Diekelmann, Wilhelm & Born, 2009). There is evidence to suggest that during SWS the newly acquired information is replayed, a process which is initially mediated by the hippocampus (Lee & Wilson, 2002). The hippocampal replay of novel information facilitated by SWS boosts the consolidation of new semantic memory, by allowing the novel information to become less reliant on the hippocampus, and more reliant on neocortical areas (Wamsley & Stickgold, 2011).

For example, Rasch, Buchel, Gais and Born (2007) investigated how learning new declarative information is related to sleep parameters, using object-location pairings. During learning, the locations were also paired with specific odors, in order to cue the participants. While the participants slept overnight in the lab, the same odors were presented to the participants in order to reactivate their memory for the learned location pairings. It was found that participants' memory for the location pairing was benefited only by cued "reactivation" of location memory by presenting the odors during SWS, but not during REM sleep or wakefulness. This benefit of cued "reactivation" during SWS was also correlated with increased hippocampal activation, as evidenced by fMRI.

Rasch et al.'s (2007) findings are important for the predictions of this study, given that it can be hypothesized based on other fMRI studies (e.g. Takashima et al. 2006; 2009) and on the basis of the predictions of the CLS models (e.g. McClleland et al., 1995) that more arbitrary mappings (compared to systematic) are more hippocampally dependent, and hence arbitrary mapping consolidation is benefited by sleep. As there is some evidence to suggest that the consolidation of more hippocampally dependent information is mediated by SWS (e.g. Rasch et al., 2007), it may be predicted that the consolidation of arbitrary mappings in language should also be mediated by SWS. Indeed, in line with the proposed SWS-mediated hippocampal reactivation mechanisms, some previous studies have found positive correlations between word-pair recall and SWS (e.g. Plihal & Born, 1997). Consequently, on the basis of evidence linking SWS with declarative recall, explicit recall tasks (such as the one we have been using) should also be mediated by SWS.

However, there is contrasting evidence for the role of SWS from Payne et al. (2009) using the Deese-Roediger-McDermott (DRM) paradigm, which has a semantic processing component. Participants learned 8 (spoken) DRM lists. Each DRM list consisted of 12 veridical words that were semantically associated with a “critical” word, such as “window”. The critical word itself was not heard. Participants recalled the words after overnight sleep or a period of daytime wakefulness. Behaviourally, it was found that participants in the sleep group, in comparison to the wake group, had higher recall of the critical words (e.g. “window”, which was not heard), but not of the veridical, trained words. Memory for the veridical, trained words was lower following both wake and sleep. The functional purpose of sleep benefiting critical (i.e. essentially false) word recall, may be that it reflects the extraction of the semantic gist of the word list. The general, semantic information may be deemed most relevant for future use. This extraction may be functionally related to our generalisation and extraction of phonological - semantic gender mappings within our Experiments, although we did not see a benefit of sleep on the two generalisation tasks in our Experiment 1. However, in the analysis of suffix errors in the recall task also in Experiment 1 we did see some evidence to indicate that the sleep group (but not the wake group) may have abstracted the suffix-gender regularities.

In terms of correlations with sleep parameters in Payne et al.’s (2009) study, participants’ recall of trained DRM words correlated negatively (significantly) with SWS. Payne and colleagues suggested that this may be because increased SWS benefits episodic memory, whereas the information learned in the study required semantic processing. SWS may negatively influence the subsequent recall of individual item information which relies on semantic processing. This study is relevant for our Experiment 2 (i.e. this Experiment), given that our recall task (as well as our two generalisation tasks) also has a semantic processing component.

The role of SWS on the consolidation of language mappings is particularly interesting to explore using our stimuli, given that we have both arbitrary and systematic mappings in the same input. As already mentioned, on the basis of the CLS’s (e.g. McClelland et al., 1995) suggestions that systematic mappings are less hipocampally-dependent (compared to arbitrary mappings), we would not expect a benefit of sleep (and SWS specifically) for the extraction of regularities. And yet, as reviewed in Chapter 1, there is previous evidence which suggests that SWS might play a role in the extraction of regularities contained within systematic mappings. Specifically, previous studies have found that increased SWS was related to better relational memory (Lau, Tucker & Fishbein, 2010), statistical probability abstraction (Durrant, Taylor, Cairney & Lewis,

2011) and schema abstraction (Durrant, Cairney & Lewis, 2012). Consequently, it is of interest whether SWS will be related to the extraction of regularities in Experiment 2 (i.e. in this Experiment) or not.

To summarize, there is some evidence from the presented studies to suggest that SWS may mediate the strengthening of hippocampally–dependent declarative information (e.g. Rasch et al., 2007). This is important, given that we are expecting (on the basis of the CLS, McClelland et al., 1995) that our arbitrary mappings are more hippocampally-dependent than our systematic mappings. However, evidence from Payne et al. (2009) suggests that SWS may be *negatively* related to individual item veridical recall which relies on semantic processing. Moreover, previous evidence also indicates that SWS may facilitate the extraction of regularities (e.g. Durrant et al., 2011; Durrant et al., 2012; Lau et al., 2010).

3.2 Experiment 2

Aims and predictions

Based on the findings from previous research, Experiment 2 aimed to investigate whether SWS can be correlated with the sleep-associated consolidation of arbitrary individual mappings which was found in Experiment 1. In Experiment 1, participants were allocated to either the daytime wake or overnight sleep group, with participants in the sleep group going home to sleep. Experiment 2 consisted of one group of participants who slept in the laboratory bedroom overnight, so that polysomnography could be used in order to measure sleep parameters, such that these sleep parameters could be correlated with behavioural measures.

Firstly, we are expecting to replicate the behavioural benefits for arbitrary but not systematic mappings found in Experiment 1. Based on the results of the previously mentioned studies, such as Rasch et al. (2007) suggesting that sleep benefits the more hippocampally-dependent mappings, and on the predictions of the CLS models (e.g. McClleland et al., 1995) that arbitrary mappings should be more hippocampally-dependent, it was hypothesized that SWS would correlate with sleep benefits for individual arbitrary mappings (recall of individual stems and suffixes). We did not expect SWS to correlate with test measures of extraction (especially given that there was no benefit of sleep for extraction of neither determiners nor suffixes as measured by the generalisation tasks in Experiment 1). Also, based on the results of Experiment 1, it was expected that behaviourally, there would be evidence of extraction of the determiner-gender regularity mapping. Given the mixed evidence for suffix-gender regularity extraction in Experiment 1, this study provides a further opportunity to explore the learning of the suffix mappings.

3.2.1 Method

Experiment 2 used the same design and procedure as Experiment 1 (i.e. the same set up for the 2 sessions, with the same tasks), but Experiment 2 was an *overnight* sleep study recording polysomnography. This means that unlike Experiment 1, participants in Experiments stayed in the lab between session 1 and session 2.

Participants

Seventeen native English participants from the University of York participated in this study. All of the participants were right-handed and female. Participants' mean age was 20.53 ± 1.50 [SD]. The data from all 17 participants were included in the behavioural analyses. One participant's polysomnography data was incomplete (due removal of the bilateral electrodes during the night), and therefore data from sixteen participants is included in the correlations with sleep stages.

Using a pre-study questionnaire, participants reported not having any language, sleep, psychiatric, neurological or hearing disorders, and were not on any psycho-active medication, had no history of drug or alcohol abuse, did not have any skin conditions, had not been unwell in the past 7 days, and had not had any problems sleeping in the last 2 weeks. As with Experiment 1 participant criteria, participants were asked to refrain from caffeinated products (in drinks or food), smoking, alcohol, drugs and daytime naps in the 24 hours prior to starting the first session and until both sessions were completed. Participants were instructed to sleep for a minimum of 6 hours the night prior to the study, to have woken up before 8 am on the day of the study and to have eaten before the session.

As with Experiment 1, before starting each session, participants completed the Stanford Sleepiness Scale and a reaction time vigilance test to check alertness levels. The vigilance measures (both the RT task and the questionnaires) are described and presented in the Appendix. Participants received £30 in return for participation and all gave their written consent.

Materials

The materials used were identical to those used in Experiment 1 (for details see Chapter 2 Method, Materials section).

Procedure

All of the participants were in the overnight sleep group. On nights where 2 participants were tested simultaneously, their arrival time was staggered, such that the average arrival time was 20.10. On nights where 1 participant was tested, the arrival time was at 20.30. As a reminder, participants were trained when they arrived (i.e. session 1 tasks summarized in Table 5 took place upon arrival). Participants were tested in the lab the following morning (i.e. the session 2 tasks summarized in Table 5 took place after overnight sleep in the lab, meaning that participants were in the lab throughout the Experiment). The order of tasks was exactly the same as in Experiment 1 and is illustrated in Table 5 as a reminder.

Table 5. Procedure (Order of tasks).

Session1

Tasks	Stimuli
Vigilance	Vigilance task and Questionnaires (including Stanford Sleepiness Scale)
Declarative	2D Object location exposure and training
Procedural	Finger-tapping exposure
Language Training	Repetition task
	Word-picture matching
Procedural	Finger-tapping training

Session 2

Tasks	Stimuli
Vigilance	Vigilance task and Questionnaires (including Stanford Sleepiness Scale)
Declarative	2D object location delayed test
Procedural	Finger-tapping retest
	Finger-tapping generalization*
Language Tests	Recall
	Determiner selection
	Suffix & determiner and suffix generalization**
	Translation recognition

*New sequence. **New items. All other tasks contained trained items only.

Upon arrival to the sleep laboratory, the polysomnography set up procedure was explained. Next, the participant signed consent forms, filled in the demographics, sleep habits and Epworth questionnaires and got changed into their sleeping clothes. Next, the participant was set up for polysomnography recording (see details below).

After this, the participant filled in the Stanford Sleepiness Scale and completed the session 1 tasks (which took approximately 50 minutes).

Following completion of the tasks and after a final polysomnography signal impedance check, the lights in the participant's laboratory bedroom were turned off (across the 17 participants, lights were turned off at 23.10 on average, ranging from 22.15 to 23.30). The participant was woken up 8 hours following lights out and given at least 30 minutes (40 minutes on average) to get ready (e.g. have a shower and something to eat). This 30-minute period was given to allow for the effect of sleep inertia to dissipate. The participant then completed the session 2 tasks (this took 50 minutes on average).

Polysomnography (PSG) set up.

In accordance with the American Academy of Sleep Science Manual (AASM, 2007), the International 10-20 system was used for polysomnography setup (Danker-Hopfe et al., 2009). All filter settings, sampling rates and impedance levels were also in accordance with the specifications of the AASM (Iber, Ancoli-Israel, Quan, 2007). Data was recorded from 13 channels: 6 of the scalp electrodes (applied bilaterally) were referenced against the contra-lateral mastoids (F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1). The left (LOC) and right (ROC) electro-oculographic (EOG) channels were used to record eye movements. The X2, X3 and X4 electromyographic (EMG) channels were used to record chin movements. An actigraph was also fixed around the participant's waist overnight in order to measure body movement. The Embla N7000 headbox system and Remlogic software were used for recording.

Scoring

Data was scored manually using Remlogic in accordance with the AASM guidelines, in 30 second epochs. Data was separated into 4 sleep stages: stage 1 (N1), stage 2 (N2), slow wave sleep (N3) and rapid eye movement (REM) sleep.

3.2.2 Results

For each task, the current results (i.e. of Experiment 2) will be presented first. A comparison of behavioural performance between Experiment 2 and Experiment 1's PM-AM sleep group will be presented second. The sleep manipulation was more controlled in Experiment 2, but the expectation was that the behavioural performance should be comparable between the two sleep groups.

Behavioural Results

Language tasks

Training task (session one)

As with Experiment 1, participants were trained on the artificial language using a word-picture matching task. Also in line with Experiment 1, two languages were used in order to counterbalance the stimuli (i.e. whether *"ked"* mapped on to masculine versus feminine natural gender of pictures). There were 9 participants in language version 1 and 8 participants in language version 2. As with Experiment 1, both accuracy and RT measures were taken into account in this task to provide an indication of the learning of the language.

Accuracy analysis

Firstly, an independent samples t-test was run with the independent variable of language version (version 1 / version 2). Accuracy at the end of training was the dependent variable. There were no significant differences between language versions ($t(15) = .28$, $p = .784$). As expected, participants exposed to either language version reached a similar standard of language knowledge by the end of the training task (see Figure 12).

Performance at the end of training was compared with performance on Experiment 1.

A two way between-subjects ANOVA was run with the between-subjects variables of Experiment (Experiment 1 sleep group, Experiment 2 sleep group) and language version (1 or 2). The dependent variable was accuracy (proportion correct). There was no significant main effect of group ($F(1, 40) = 2.49$, $p = .360$), no significant main effect of language version ($F(1, 40) = 3.47$, $p = .314$), and no significant interaction between group and language version ($F(1, 40) = .21$, $p = .652$). This suggests that behaviourally, training performance was comparable between the two Experiments, as shown in Figure 12.

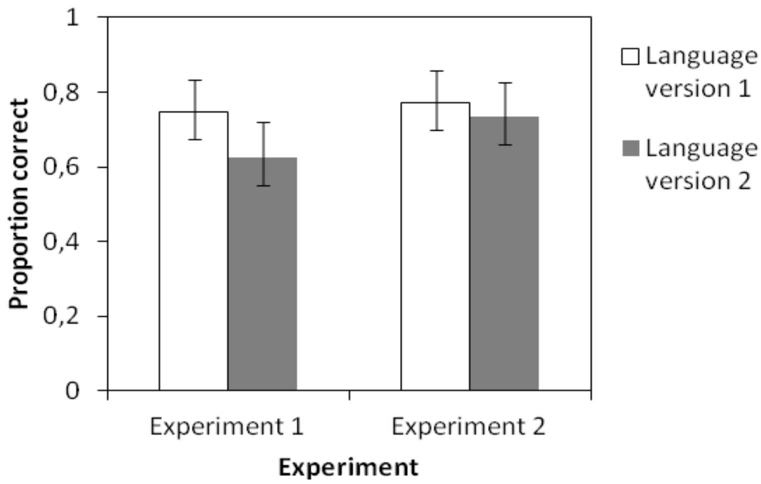


Figure 12. Performance at the end of training.

Correct RTs analysis on the training task

As with Experiment 1, the RT data for correct responses was analysed separately for target trials and separately for noise trials. Correct responses to target trials were hits, and correct responses to noise trials were correct rejections. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), and the between subjects variable of language (version 1 or version 2). First, the dependent variable was the change in target RTs compared to block 1. One participant did not have any correct responses on the target RTs, and therefore had to be excluded from this analysis.

There was a significant main effect of block ($F(1, 14) = 12.76$, $p < .001$), no significant main effect of language ($F(1, 14) = .88$, $p = .365$), and no significant interaction between block and language ($F(1, 14) = 1.45$, $p = .241$). There was no evidence to suggest that participants' performance differed across the two language versions. In addition, as seen in Figure 13, all participants improved across the blocks, indicating that participants' knowledge of the language improved across the exposures.

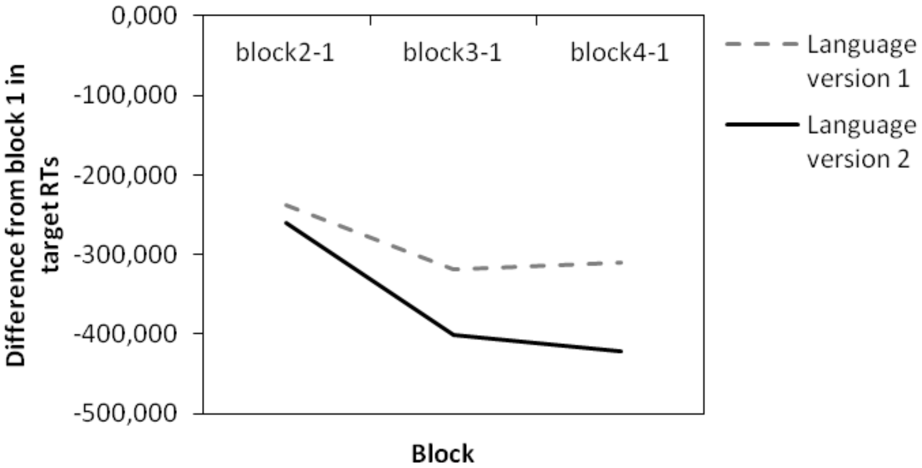


Figure 13. Difference from block in target RTs in Experiment 2.

Secondly, the dependent variable was the change in noise RTs compared to block 1. There was a marginally significant main effect of block ($F(1, 15) = 2.77$, $p = .079$), no significant main effect of language ($F(1, 15) = 1.05$, $p = .322$), and no significant interaction between block and language ($F(1, 15) = 1.35$, $p = .274$). As seen in Figure 14, there was a marginal improvement in noise RTs across blocks.

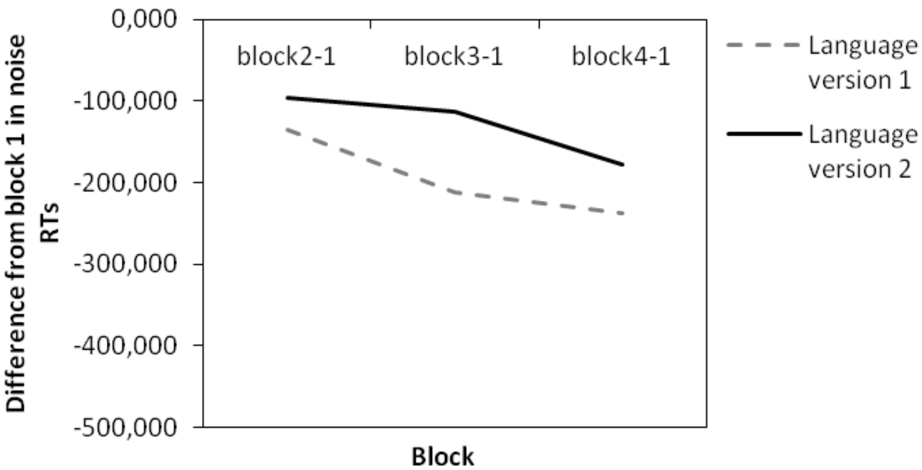


Figure 14. Difference from block 1 in noise RTs in Experiment 2.

Overall, all participants in Experiment 2 improved across the additional exposures to the language, indicating learning of the language, regardless of the language version. Next, the training task RT data from Experiment 2 was compared with the training task RT data from Experiment 1 for target RTs separately from the noise RTs, as before.

The target RT data from Experiment 2 was compared with RT data from Experiment 1.A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), and the between subjects variables of language (version 1 or version 2) and Experiment (Experiment 1/ Experiment 2). The dependent variable was the change in target RTs compared to block 1. There was no significant main effect of Experiment (F(1, 38) = .00, p = .960), a significant main effect of block (F(1, 38) = 27.89, p < .001), and no significant Experiment x block interaction (F(1, 38) = .38, p = .684). This suggests that participants from both Experiments improved a similar amount across the blocks, indicative of similar levels of learning, as shown in Figure 15.

In addition, there was a significant main effect of language (F(1, 38) = 4.35, p = .044), and a marginally significant interaction between block and language (F(1, 38) = 2.71, p = .073). There was no significant interaction between language and Experiment (F(1, 38) = .42, p = .520), and no significant three way interaction between block, language and Experiment (F(1, 38) = .05, p = .951). As shown in Figure 15, participants' improvement was greater in language version 2 (Figure 15b) than in language version 1 (Figure 15a).

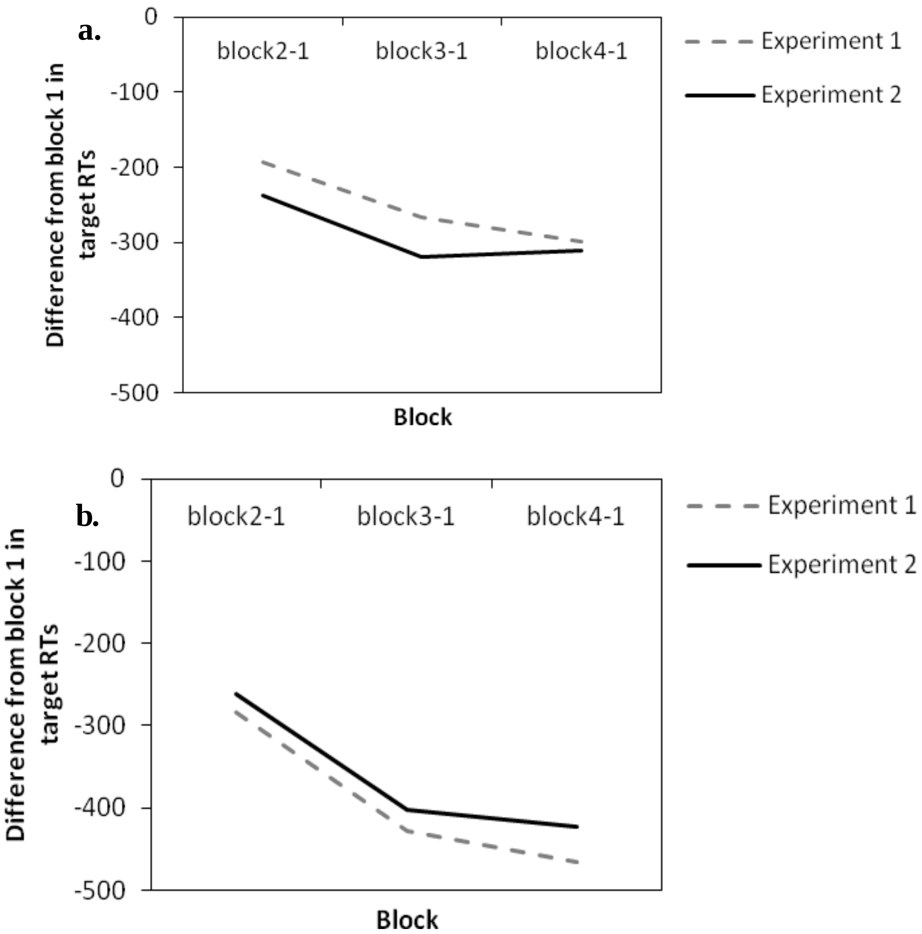


Figure 15. Difference from block 1 in target RTs (Figure 15a. shows language version 1 RTs, Figure 15b. shows language version 2 RTs).

In addition, the noise RT data from Experiment 2 was compared with RT data from Experiment 1. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), and the between subjects variables of language (version 1 or version 2) and Experiment (Experiment 1/ Experiment 2). First, the dependent variable was the change in noise RTs compared to block 1.

There was a marginally significant main effect of Experiment ($F(1, 38) = 3.67$, $p = .063$), a significant main effect of block ($F(1, 38) = 13.20$, $p < .001$), and no significant Experiment x block interaction ($F(1, 38) = 1.40$, $p = .253$). This suggests that all participants improved across the blocks. In addition, participants from Experiment 2 improved less across the blocks, than participants from Experiment 1, as shown in Figure 16. Crucially, these are noise RTs (i.e. correct rejections).

In addition, there was no significant main effect of language ($F(1, 38) = .00$, $p = .973$), no significant interaction between block and language ($F(1, 38) = .39$, $p = .678$).

There was a marginally significant interaction between language and Experiment ($F(1, 38) = 3.68$, $p = .063$), and no significant three way interaction between block, language and Experiment ($F(1, 38) = 1.47$, $p = .237$).

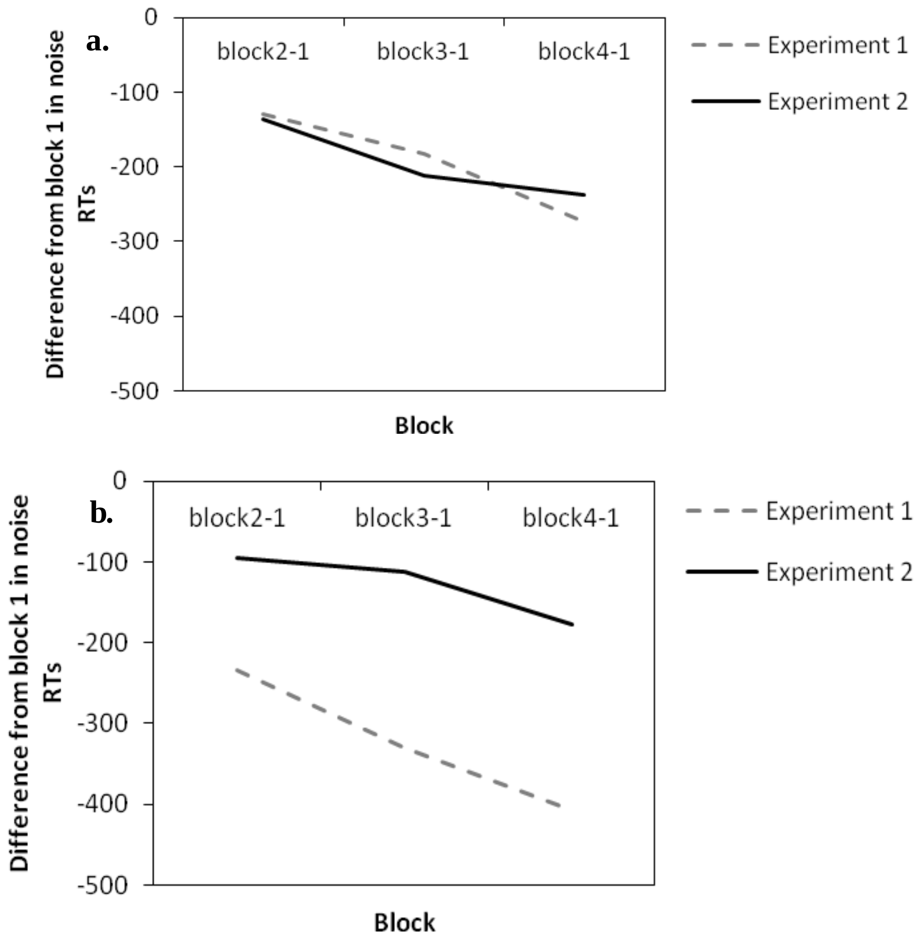


Figure 16. Difference from block 1 in noise RTs (Figure 5a. shows language version 1 RTs, Figure 5b. shows language version 2 RTs).

Overall, the accuracy analysis indicated that all of the participants had overall similar levels of knowledge at the end of training (regardless of the language version). The results from the RT analysis of the training task suggested that all of the participants improved across the blocks, which is reflective of learning of the artificial language. Although the analysis of the target correct RTs across the two Experiments indicated that improvement was overall greater in language version 2 than in language version 1 (indicative of greater learning in language version 2), given that the focus is on comparing Experiments 1's sleep group with Experiment 2's sleep group, the analyses of the session two tasks will not take language into account.

Language tests (Session 2)

Most arbitrary mapping: vocabulary

Translation Recognition Task

As with Experiment 1, this task was used to test participants' memory of the trained items (vocabulary). Also in line with Experiment 1, the non-parametric discrimination (A') formula (Donaldson, 1992) was also used to account for participant's discrimination of item-pairs which matched and those which did not. The average A' discrimination across the 17 participants was 0.85, indicating a good level of discrimination between matching and mismatching pairs.

A between subjects t-test was run to compare performance with Experiment 1. The between subjects variable was Experiment (Experiment 1 sleep group, Experiment 2 sleep group). The dependent variable was A' discrimination, in order to account for discrimination between match and mismatch trials. There were no significant differences in performance between Experiment 1 sleep group and Experiment 2 sleep group ($t(42) = .88$, $p = .384$). This suggests that recognition memory of the arbitrary mappings in Experiment 2 was comparable with recognition memory of the arbitrary mappings in Experiment 1, as shown in Figure 17.

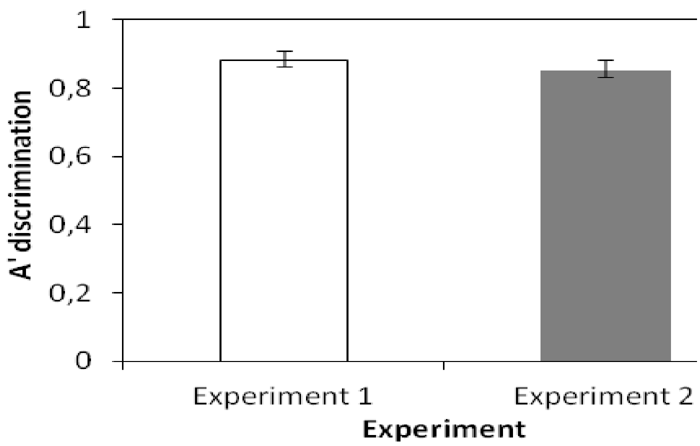


Figure 17. Performance on the translation recognition task.

Recall (stems)

As with Experiment 1, this task was used to assess participants' memory of the trained individual words. Each picture was presented twice – once the picture on its own, the second time with the first grapheme of the corresponding word written underneath as an additional cue. Also in line with Experiment 1, we analysed separately the performance on the stem and the suffix portions of the words, out of total productions. This was done in order to separate the more arbitrary (stems) from the more systematic (suffixes) mapping. The average proportion of stems recalled correctly in Experiment 1 was .088 when both the picture and grapheme were present, and .073 when the picture only was present. The average proportion of stems recalled correctly in Experiment 2 was .257 when both the picture and grapheme were present, and .179 when the picture only was present. This indicates that recall rates in Experiment 2 may have been slightly higher than in Experiment 1.

To analyse stem recall in Experiment 2, a paired samples t-test was run with the independent variable of cue type (picture only/ picture and grapheme) and the dependent variable of proportion correct (out of total productions). There was a significant difference between the two cue types ($t(16) = 4.19$, $p < .001$). Participants recalled significantly more stems correctly, when the additional grapheme cue was also present, than when the picture only was present, as in Experiment 1.

Stem recall in Experiment 2 was compared with stem recall in Experiment 1. A mixed ANOVA was run with the between-subjects variable of Experiment (Experiment 1 sleep group, Experiment 2 sleep group), and the within subjects variable of cue type (picture and grapheme, or picture only). The dependent variable was proportion correct (out of total productions). There was a significant main effect of Experiment ($F(1, 42) = 11.16$, $p < .001$), a significant main effect of cue type ($F(1, 42) = 22.16$, $p < .001$), and a significant interaction between Experiment and cue type ($F(1, 42) = 9.12$, $p = .004$).

As shown in Figure 18, participants in Experiment 2 had overall higher stem recall accuracy than participants in Experiment 1. In addition, all participants' stem recall was benefited by the presence of the additional grapheme cue, but this benefit was greater in Experiment 2 than in Experiment 1. Aside from the smaller number of participants in Experiment 2, a possible explanation for the better stem recall in Experiment 2 may be that the overnight sleep in the lab was more controlled, which may have benefited consolidation. Indeed, as shown in the appendix, participants in Experiment 2 were better rested than participants in Experiment 1, according to vigilance measures.

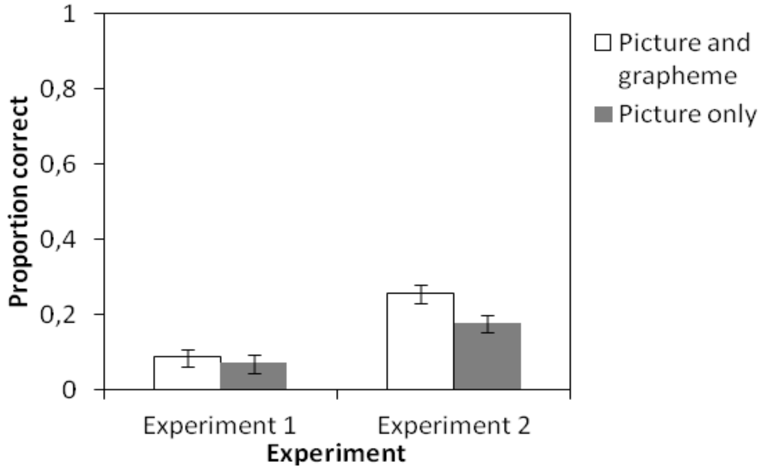


Figure 18. Stem recall accuracy.

To summarize the findings from the vocabulary tests, participants had overall a good level of recognition memory, and the performance was comparable to Experiment 1. Stem recall was benefited by the presence of the first grapheme cue. Stem recall was also higher in Experiment 2, than in Experiment 1, potentially because of the more controlled sleeping conditions.

Intermediate systematicity mapping: suffixes (grammar)

Recall (suffixes)

The recall accuracy of the suffixes (which were separated from stems in the transcriptions) was analysed using a paired sample t-test. The independent variable was cue type (picture only/ picture and grapheme) and the dependent variable was proportion correct (out of total productions). There was a significant difference between the two cue types ($t(16) = 3.34$, $p = .004$). Participants recalled significantly more suffixes correctly, when the additional grapheme cue was also present, than when the picture only was present.

Suffix recall in Experiment 2 was compared with suffix recall in Experiment 1. A mixed ANOVA was run with the between-subjects variable of Experiment (Experiment 1 sleep group, Experiment 2 sleep group), and the within subjects variable of cue type (picture and grapheme, or picture only). The dependent variable was proportion correct (out of total productions). There was a marginally significant main effect of Experiment ($F(1, 42) = 3.47$, $p = .069$), a significant main effect of cue type ($F(1, 42) = 25.70$, $p < .001$), and no significant interaction between Experiment and cue type ($F(1, 42) = .28$, $p = .599$). As shown in Figure 19, and as indicated by the marginally significant effect of Experiment, participants in Experiment 1 had somewhat higher suffix recall than participants in Experiment 2. All participants regardless of Experiment had higher recall when the grapheme cue was additionally present.

In addition, it may be important to mention that stem recall was higher in Experiment 2 than in Experiment 1. It was the opposite for suffix recall, which was marginally better in Experiment 1, than Experiment 2. As already mentioned, these differences between Experiments may be explained by the smaller number of participants in Experiment 2. Although originally suggested as an explanation for stem recall differences between Experiments, it is unlikely that the more controlled sleep conditions in Experiment 2 and greater vigilance in Experiment 2 (as shown in the Appendix) can explain the lower suffix recall in Experiment 2. An alternative explanation may be that there were individual differences in the participants between the two Experiments.

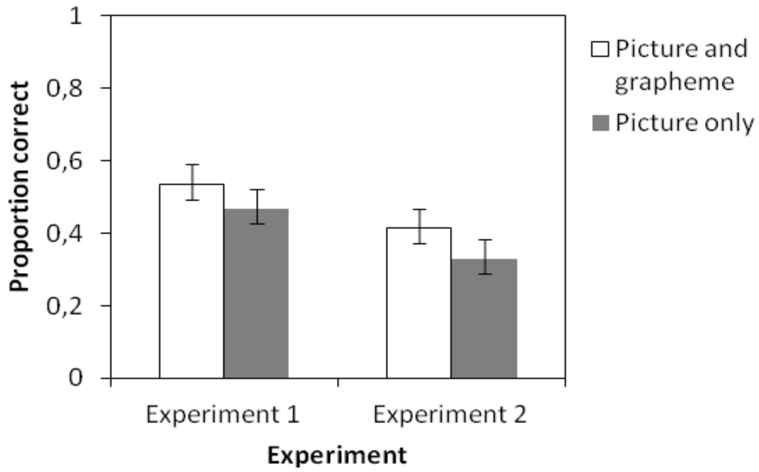


Figure 19. Suffix recall accuracy.

Analysis of errors on the suffix recall task

As with Experiment 1, suffix errors reflecting correct suffix-gender mappings (proportion of correct gender suffixes out of the incorrect but trained suffixes) for both picture only and picture and grapheme cue types were compared against the chance level of 0.33, as each produced word could have three possible incorrect trained suffixes and only one of the three could be of the correct gender. Abstraction of the suffix to picture gender mapping would be reflected by correct suffix-gender mappings being produced above chance.

Two participants in the picture and grapheme cue type and four participants in the picture only cue type were excluded from the analysis, because none of their incorrect responses included one of the trained suffixes. Thus, the analysis was done on fifteen participants in the picture and grapheme cue type and thirteen participants in the picture only cue type.

One sample t-tests against chance revealed that the proportion of suffix errors reflecting correct suffix-gender mappings was *not* significantly higher than chance for either of the cue types (see Table 6 for results of the one sample t-tests). These findings suggest that participants were relying on their knowledge of the *individual* suffixes, as opposed to on their abstraction of the suffix-gender regularities to complete the suffix recall task.

Table 6. Results of correct suffix-gender errors analysis in the recall task using one sample t-tests.

	t	df value	*p*-value	Mean proportion
Picture and grapheme (N=15)	.74	14	.472	.383
Picture only (N=13)	.78	12	.450	.407

Suffix generalisation task

As with Experiment 1, this task was used to test participants' extraction of the suffix-gender regularity. The analysis of suffix errors on the recall task suggests that participants were relying on their knowledge of the *individual* suffixes to complete the recall task. Consequently, it was of interest whether this task would provide evidence that the suffix-gender regularities had been extracted. A paired samples t-test was run with the independent variable of consistency (items consistent vs inconsistent with the trained regularities) and the dependent variable of proportion endorsed. There were no significant differences between consistency types ($t(16) = -1.44$, $p = .170$), as shown in Figure 20.

A one sample t-test was run on the consistent items, and separately on the inconsistent items, to compare rates of endorsement of consistent and inconsistent items against the chance level of 0.5. Participants endorsed consistent items significantly above chance ($t\ (16) = 4.40$, $p < .001$), and inconsistent items marginally above chance ($t\ (16) =$

2.05, $p = .058$). As there was no main effect of consistency there was no strong evidence of suffix-gender regularity extraction, as with Experiment 1. The above chance performance for consistent items and close to above chance for inconsistent was not completely in line with the pattern expected had the suffixes been extracted. This is because full extraction of suffix-gender regularities would be reflected by significantly below chance performance for the inconsistent items.

Nonetheless, the above chance performance suggests that something may have been extracted, for instance the determiner-gender regularity. As a reminder of the items in this task (previously described in Chapter 2 Method, Materials section), for the inconsistent items, the *suffix* was inconsistent with the determiner and the natural gender of the previously unseen pictured character. This means that in the suffix generalisation task, the determiner-gender regularity was always consistent with the trained determiner-gender regularity, even on the items where the suffix-gender regularity was inconsistent.

The endorsement of the consistent and inconsistent items on the suffix generalisation task in Experiment 2 was compared with endorsement in Experiment 1. A mixed ANOVA was run with the between-subjects variable of Experiment (Experiment 1 sleep group, Experiment 2 sleep group), and the within subjects variable of consistency (items consistent vs inconsistent with the trained regularities). The dependent variable was proportion endorsed. There was no significant main effect of Experiment ($F(1, 42) = .35$, $p = .560$), there was a marginally significant effect of consistency ($F(1, 42) = 3.26$, $p = .087$), and no significant interaction between Experiment and consistency ($F(1, 42) = 2.18$, $p = .148$).

As shown in Figure 20, based on the pattern of endorsement, and the marginal main effect of consistency when the data from both Experiments was combined, there was *some* indication that *something* may have been extracted, but no *clear* evidence to suggest that the suffix-gender regularity was extracted. However, there is some indication that the pattern of endorsement was *not* due to the extraction of the suffix-gender regularity, as the higher than chance endorsement for inconsistent items is not in line with the pattern expected had the suffix-gender regularity been extracted. Consequently, there is *some* indication in both Experiments that participants may have been relying on the determiner-gender regularity extraction instead, to complete this task.

Overall, behaviourally the findings from this task may be comparable across the two Experiments, as there was no main effect of group, and no interaction between group and consistency. This comparability of performance is a key question in the analyses, as it suggests that behaviourally performance across Experiments 1 and 2 is comparable. Consequently, if any correlations between sleep parameters and the suffix generalization task are found in Experiment 2, they could in principle generalize to Experiment 1, on the basis of similar behavioural performance.

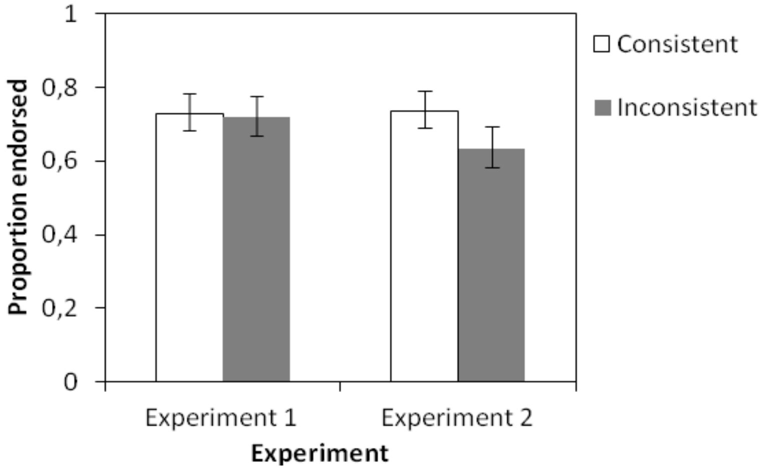


Figure 20. Proportion endorsed on the suffix generalisation task.

To summarize the findings from the suffix tests, suffix accuracy was overall higher when the grapheme cue was present (than when the picture only was present). There was no *strong* evidence of suffix-gender regularity extraction. Yet, there was *some* indication on the basis of above chance performance on the suffix generalization task, and the main effect of consistency only when the data from the two Experiments were combined that *something* had been extracted. Specifically, participants were potentially relying on their knowledge of determiner-gender regularities to complete this task. In line with this, the suffix-error analysis on the recall task has suggested that participants were relying on their knowledge of the *individual* suffix-picture mappings, as opposed to the suffix-gender regularities, when recalling suffixes. The analyses of the determiner tests in the next section will confirm whether participants may indeed have relied on their knowledge of the determiner-gender regularity, as we have suggested.

Most systematic mapping: determiners (grammar)

Determiner selection task

As with Experiment 1, participants' memory for the trained determiner-picture mappings was assessed using a determiner selection task. A one sample t-test was run to compare the accuracy (proportion correct) against the chance level of 0.5. Participants' performance on the determiner selection task was significantly above chance (t (16) = 3.79, p = .002), suggesting that participants had memory traces for the determiner-picture mappings.

A between subjects t-test was run to compare performance with Experiment 1. The between subjects variable was Experiment (Experiment 1 sleep group, Experiment 2 sleep group). The dependent variable was proportion correct. There were no significant differences between groups (t(42) = .62, p = .685). This suggests that the memory for the individual determiner-picture mappings in Experiment 2 was comparable with Experiment 1, as shown in Figure 21.

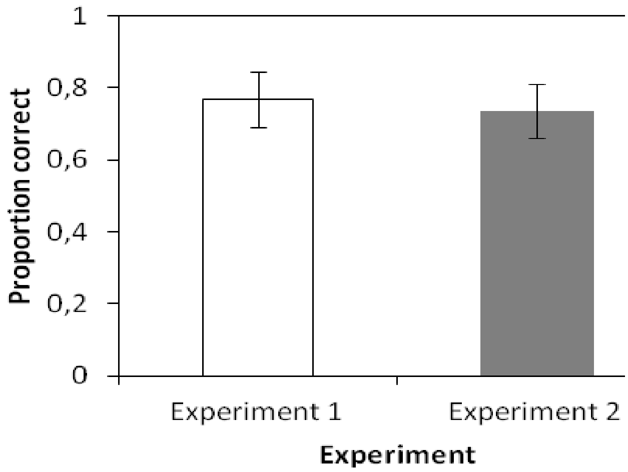


Figure 21. Performance on the determiner selection task.

Determiner and suffix generalisation task

As with Experiment 1, this task was used to test participants' generalization of the determiner-gender regularity. A paired samples t-test was run with the independent variable of consistency (consistent/ inconsistent items) and the dependent variable of proportion endorsed. There was a significant difference between consistency types (t(16) = 2.62, p = .019), as shown in Figure 22.

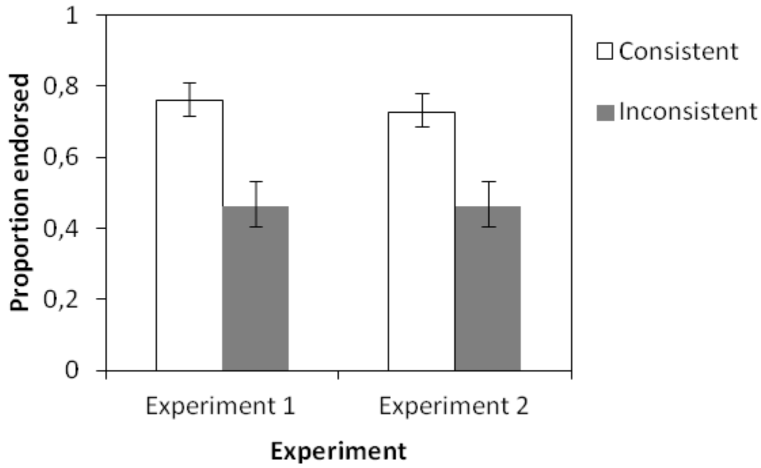


Figure 22. Proportion endorsed on the determiner and suffix generalisation task.

As with Experiment 1, a one sample t-test was run on the consistent items, and separately on the inconsistent items, to compare rates of endorsement of consistent and inconsistent items against the chance level of 0.5. Participants endorsed consistent items significantly above chance (t (16) = 4.41, $p < .001$), and inconsistent items at chance (t (16) = -.477, $p = .640$). Based on the main effect of consistency, there was evidence to suggest that the determiner-gender regularity had been extracted. In addition, the above chance performance is also in line with the pattern expected if the determiners had been extracted. This evidence of determiner-gender regularity extraction is in line with findings from Experiment 1. Consequently (as with the suffix generalisation task), if any correlations between sleep parameters and the determiner and suffix generalisation task are found in Experiment 2, they could in principle generalize to Experiment 1, on the basis of similar behavioural performance.

The endorsement of the consistent and inconsistent items on the determiner and suffix generalisation task in Experiment 2 was compared with endorsement in Experiment 1. A mixed ANOVA was run with the between-subjects variable of Experiment (Experiment 1 sleep group, Experiment 2 sleep group), and the within subjects variable of consistency (items consistent vs inconsistent with the trained regularities). The dependent variable was proportion endorsed. There was no significant main effect of Experiment ($F(1, 42) = .09$, $p = .755$), there was a significant effect of consistency ($F(1, 42) = 20.12$, $p < .001$), and no significant interaction between Experiment and consistency ($F(1, 42) = .06$, $p = .804$). This provides evidence of determiner-gender regularity extraction in both Experiments 1 and 2, and suggests that behaviourally performance on this task was comparable across the two Experiments, as shown in Figure 22. Consequently, if any correlations are found in Experiment 2 between SWS and the determiner and suffix generalisation task, it may be reasonable to suggest that the same correlation might have applied to the behavioural findings of Experiment 1.

To summarize, the findings from the determiners tests suggest that participants had both memory of the individual determiner-picture mappings, and extraction of the determiner-gender regularity. This suggests that the above chance performance on the suffix generalisation task may have been driven by the participants' knowledge of determiner-gender regularities, as opposed to suffix-gender regularities. In addition, individual memory for determiner-picture mappings and extraction of the determiner-gender regularities was behaviourally comparable between Experiments 1 and 2.

Participants in Experiment 2 recalled more stems correctly, and fewer suffixes correctly in comparison to participants from Experiment 1. Given that the performance on stem recall is further in the direction of a sleep benefit, it may be that the behavioural performance of the two groups is comparable on stem recall. However, given the lower suffix recall in Experiment 2, it is unclear whether suffix recall performance is entirely comparable.

Correlations with sleep data

The average sleep parameters are presented in Table 7, and these average suggest that participants had a typical sleep architecture. Specifically, according to the AASM manual for the scoring of sleep (2007), typical sleep architecture includes most time spent in stage 2 (approximately 50% is typical), less so in SWS (approximately 25% is typical), followed by stage 1, with the least time spent in REM. Consequently, in Experiment 1 the proportion of time spent in each sleep stage was typical. We analysed our data on the basis of time spent in SWS as a percentage of total sleep time.

The test measures of interest for correlations with sleep stages were the stem recall and suffix recall, as the measures of memory of the arbitrary mappings. To clarify, the analysis of the suffix errors on the recall task has suggested that participants were relying on their memory of individual suffixes, as opposed to suffix-gender regularities to complete the suffix recall task. This suggests that the suffix recall task reflects the memory of the arbitrary mappings, i.e. the memory for the specific picture-suffix mappings. In addition, we were interested in correlating the sleep stages with performance on the 2 generalisation tasks, as they were the measures testing extraction of firstly suffixes and secondly determiners. These 4 measures allowed to investigate the role of sleep (specific sleep stages) in consolidation of arbitrary mapping (stem and suffix recall) and extraction of regularities (the 2 generalisation tasks).

Table 7. Sleep parameters for participants. Parentheses denote the standard deviation.

Sleep parameter	Mean time in minutes	Time as a percentage of total sleep time
Total sleep time	429.5 (46.01)	
Wake time after sleep onset*	35.99 (33.43)	
Sleep latency	12.13 (11.87)	
Stage 1	93.45 (24.09)	21.74 (5.14)
Stage 2	196.22 (48.05)	45.38 (8.41)
SWS	99.06 (22.6)	23.22 (5.12)
REM	40.75 (15.88)	9.63 (3.97)

*Wake time after sleep onset refers to the amount of time spent awake between initially falling asleep and becoming completely awake. E.g. if the participant fell asleep at 23.00 initially, and become completely awake at 07.00, but was also awake between 03.00 and 03.03, the wake time after sleep onset would be 3 minutes.

We were expecting SWS to correlate with measures of arbitrary mappings (i.e. stem recall and suffix recall), but not with measures of systematic mappings (the suffix generalisation task, and the determiner and suffix generalisation task). To test this hypothesis, correlation analyses were run with time spent in SWS (as a percentage of total sleep time) and stem recall, suffix recall, suffix generalisation task, and the determiner suffix generalisation task.

In the recall task (stems and suffixes), participants saw each picture twice, once on its own, and once with the first grapheme of the word present. For the purpose of the analysis presented here, the recall data was averaged across both types of cues, given that it was participants' memory for trained individual items that was of interest, rather than how their memory was affected by cue type (both cues were offered in order to improve participants' recall given that performance was low, rather than using cue type as a measure of interest).

As 6 correlations were run, the Bonferroni correction was applied, and hence the *p*- value was set at .008. As shown in Table 8, there was a significant negative correlation between stem recall and proportion of time spent in SWS (SWS %), but no significant correlation with suffix recall. There were no significant correlations between proportion of time spent in SWS (SWS%) and the measures of generalization, as expected. As shown in Table 8, there were no other significant correlations between language test measures and correlations that survived the Bonferroni correction (whereby the *p*- value was set at .008).

Table 8. Correlations between session 2 language test measures and SWS%.

Test Measure (N=16)		REM %	Stage 1 %	Stage 2 %	SWS %	Sleep spindle count*
Stem recall	*r*	-.110	-.198	.562	-.635	.208
	p	.685	.463	.024	.008	.456
Suffix recall	*r*	-.267	-.011	.370	-.388	.234
	p	.318	.969	.159	.138	.401
Translation	*r*	.252	-.361	.362	-.425	.345
	p	.347	.170	.168	.101	.208
Determiner Selection	*r*	-.223	-.115	.367	-.331	.328
	p	.407	.671	.162	.241	.232
Suffix generalization (consistent)	*r*	.146	-.481	.278	-.083	-.408
	p	.588	.059	.297	.761	.131
Suffix generalization (inconsistent)	*r*	-.089	-.225	.385	-.330	.197
	p	.744	.401	.141	.212	.482
Determiner and suffix generalization (consistent)	*r*	.065	-.363	.013	.295	-.278
	p	.811	.167	.962	.268	.316

Determiner and suffix generalization (inconsistent)	*r*	-.125	-.248	.379	-.276	.046
	p	.643	.354	.148	.301	.870

*The sleep spindle count measure was based on 15 participants, due to technical issues not permitting the analysis of 1 participant's data.

As shown in Figure 23, as the proportion of time spent in SWS increased, stem recall accuracy decreased. This was *not* in line with our expectation that the proportion of time spent in SWS would correlate *positively* with stem recall accuracy.

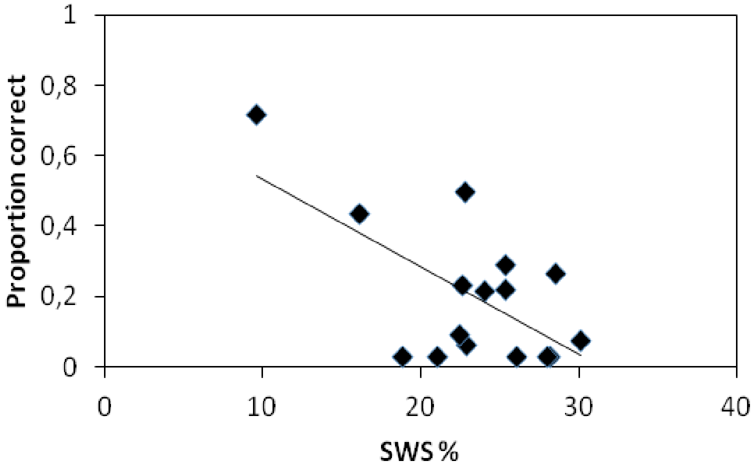


Figure 23. Stem recall accuracy (averaged across cued and non-cued items) correlation with SWS%.

3.3 Chapter Summary and Discussion

Experiment 2 added on to the behavioural findings from Experiment 1, by correlating test measures with overnight sleep parameters. As a reminder, results from Experiment 1 revealed a sleep benefit for the recall of the most arbitrary mapping (individual stems). Also in Experiment 1, regardless of sleep, there was extraction of the determiner-gender regularity mapping (as evidenced by performance on the determiner selection task, and the determiner and suffix generalisation task). Experiment 1 provided mixed evidence for the extraction of suffix-gender regularities (as evidenced by performance on the suffix recall task and suffix generalisation task), with no strong evidence of suffix-gender regularity in *either* of the groups.

In Experiment 2, as with Experiment 1, behaviourally there was evidence that the determiner-gender regularity was extracted, but no clear evidence to suggest that the suffix-gender regularity was extracted. In Experiment 2 (i.e. in this Experiment) we found that SWS correlated *negatively* with the individual, arbitrary mapping (i.e. stem recall), which was not in line with our predictions. We found no evidence of correlations between SWS and the systematic mappings (i.e. suffix-gender regularities and determiner-gender regularities).

Individual, arbitrary mappings and SWS

Based on previously found correlations of SWS with more hippocampally-dependent information (e.g. Rasch et al., 2007), it was hypothesized that performance on the stem recall task (measuring the individual, arbitrary mapping) would correlate *positively* with SWS. This is because on the basis of the CLS (e.g. McClelland et al., 1995) we are expecting the individual, arbitrary mappings to be more hippocampally-dependent than the systematic mappings. However, the findings of Experiment 2 were not in line with these predictions, as consolidation of the most arbitrary mapping (recall of individual stems) was found to correlate *negatively* with SWS, rather than positively.

Although there is little previous evidence which could predict this result, findings can be explained in reference to Payne et al.'s (2009) findings. Using the DRM paradigm, they found that SWS correlated *negatively* with memory of veridical trained items (this measure is comparable to our stem recall performance). They suggested that it may be that SWS benefits the consolidation of individual episodic memories (as previously found by Marshall and Born, 2007; Peigneux et al., 2004; Rasch et al., 2007; Takashima et al., 2007), but consequently decreases performance on tasks requiring semantic processing. The key difference would be that the retrieval of information which requires semantic processing, including word meanings (in comparison to episodic memories) is less reliant on the exact context in which the information was learned (e.g. Tulving, 2002). Our recall task, similarly to Payne et al.'s (2009) task involved semantic processing. However, our generalisation tasks would have also required semantic processing, yet we did not see a negative correlation with SWS and the two generalisation tasks. There was no evidence to suggest that increased SWS was the reason why suffix-gender regularity extraction performance (requiring semantic processing) was impaired. Consequently, the negative correlation between SWS and the individual arbitrary mapping (i.e. stems) needs further research to be fully explained.

Extraction of regularities

Both Experiments 1 and 2 have provided behavioural evidence for the extraction of the determiner-gender regularities, but not for the extraction of the suffix-gender regularities. In both Experiments, there was a lack of sleep effect for the consolidation of the systematic mappings, as we predicted on the basis of the CLS (e.g. McClelland et al., 1995). In Experiment 2, there was no indication of correlations between SWS and systematic mappings. The lack of indication for correlation between SWS and determiner-gender regularity extraction is expected, given that both Experiments 1 and 2 suggest that this extraction is *not* benefited by sleep.

Reasons why determiners were preferentially extracted over suffixes regardless of sleep will be explored in the following Chapter. The key difference between the determiners and the suffixes was that determiners were fully systematic (as there was 1 determiner per gender), and suffixes were of intermediate systematicity (as there were 2 suffixes per gender). This also meant that each determiner appeared in the input more frequently, than each suffix, because of the nature of the stimuli. This leads to the suggestion that the determiner-gender regularities were extracted, while the suffix-gender regularities were not, because determiners (relative to suffixes) were the more systematic and the more frequent mapping in the input.

This suggestion would be consistent with suggestions of the Competition model (e.g. Bates & MacWhinney, 1989). Specifically, the Competition model suggests that when two cues point to the same mapping (such as the determiners and suffixes both pointing to the same gender mapping), the cue which is higher in reliability (i.e. systematicity) and higher in availability (i.e. frequency) is assigned a higher strength in predicting the mapping. Consequently, the cue that was assigned a higher strength of predicting the mapping is extracted before the cue that was assigned a lower strength.

In the next Chapter we aim to investigate if that is indeed the case, i.e. whether the relative systematicity and frequency of cues are the key factors that determine whether the cues (i.e. regularities) are extracted or not. It is of interest to confirm whether in Experiments 1 and 2 the determiner-gender regularity was extracted (regardless of sleep) due to its higher level of systematicity and frequency in comparison to the suffix-gender regularity.

To investigate this, In Experiment 3 we are going to swap the systematicity of the mapping, making determiners of intermediate systematicity, and suffixes fully systematic.

If high systematicity and frequency are sufficient for extraction, we should see evidence of suffix-gender regularity extraction in the next Chapter, on the basis that fully systematic and more frequent determiners were extracted in Experiments 1 and 2. If this is not the case, we will we be able to explore what other factors within the language input structure may also contribute to the likelihood of extraction and generalisation.

Conclusions

In Experiment 2 we found a similar pattern of behavioural findings as in Experiment 1. Specifically, there was some evidence for the benefit of sleep for the consolidation of arbitrary mappings, but no evidence for the benefit of sleep for the extraction of regularities contained within systematic mappings. Regardless of sleep, for systematic mappings we found extraction of determiner-gender regularities, and no clear evidence for the extraction of suffix-gender regularities, consistently with Experiment 1. In addition, we found a new correlation with SWS for arbitrary mappings (i.e. individual stems), which were not in line with our predictions. Instead, the negative relationship between SWS and the consolidation of the individual, arbitrary stem mappings was in line with Payne et al.'s (2009) findings using the DRM paradigm. Overall, the experimental findings need further research to be fully understood.

In the next Chapter we seek to explore whether the degree of *systematicity* of the mapping is the only factor which determines whether regularities are extracted or not. We also continue to investigate the extent to which sleep aids the extraction of arbitrary versus systematic mappings in language learning.

Chapter 4: The ordering of information in time: the effect of swapping the systematicity of the mappings

4.1 Introduction

In Experiments 1 and 2 the factor we focused on was the degree of systematicity of the language mapping. It was found that regardless of sleep, the more systematic mappings (the determiners) were extracted, but the less systematic mappings (the suffixes) were not extracted. Experiments 3 and 4 presented in this Chapter will continue to investigate the factor of systematicity. More specifically, we are going to swap the systematicity of determiners and suffixes, in order to investigate whether systematicity is the key factor that determines whether regularities are extracted.

As suggested in the discussion of Chapter 3, the behavioural findings from Experiments 1 and 2 are consistent with suggestions of the Competition Model (e.g. Bates & MacWhinney, 1989). While there are several different components to the Competition model, only the aspects which apply to our Experiments, namely cue reliability and cue availability will be discussed. The Competition model focuses on the acquisition of form to function mappings. This is similar to what we have called form-meaning mappings within our Experiments (i.e. determiner-gender regularity mappings and suffix-gender regularity mappings). Within the model, the acquisition of form to function mappings is input-driven (as opposed to innate). Based on what is encountered in the input, each form-meaning mapping can be assigned a strength, i.e. the probability that the form predicts the corresponding meaning. This means that whether a mapping is acquired or not initially, depends on its cue strength. That is, the more strength a cue has, the more likely that mapping will be acquired. Given that in our Experiments, we are investigating the initial stages of a language's acquisition, these concepts are particularly relevant.

The key idea is that when two cues both map on to the same function (i.e. meaning), competition arises between the acquisition of the two cues. The relative strength of each cue in predicting a meaning is compared. The cue that is more likely to be acquired is the one that has relatively more strength. Although we are investigating the acquisition of cues at a word level, whilst many of the studies investigating the ideas of the Competition model have examined sentence interpretation, and in particularly the identification of the agent in the sentence, the same argument applies to our Experiments as well. For instance, it was found that Hungarian children particularly relied on case marking as a cue to identify the agent within a sentence, largely because of its higher strength, i.e. higher predictability as a cue in comparison to other cues. In other words, the cues to agent with the greatest strength were acquired first (MacWhinney, Pléh & Bates, 1985).

The strength of a cue may be determined by the simple reliability (i.e. what we have termed systematicity) and simple availability (i.e. frequency in the input) of the cue. For example, there is some evidence to suggest that initially, the simple availability of cues determines children's acquisition of cues. In scenarios where there are two cues to the same meaning, children acquire the cue that is more available first. In Hebrew, the inflectional reflexive is acquired before the periphrastic reflexive, because of its greater simple availability, i.e. because it is more present in the input (Sokolov, 1989). Consequently, cues with higher simple availability may be more likely to be acquired.

To relate the concepts of reliability and availability to our stimuli, in Experiments 1 and 2, there was 1 determiner per gender. In other words, the gender was always reliably indicated by the determiner (e.g. female characters were *always* indicated by '*tib*'). As there were 2 suffixes per gender, the gender was indicated by one suffix half of the time, and by another suffix the rest of the time (e.g. female characters were indicated *sometimes* by *'esh'*, and *sometimes* by *'eem'*). Relating this to the Competition Model, in Experiments 1 and 2, the determiners were the more reliable cue, in comparison to the suffixes.

The concept of 'cue availability' is related to the number of times that participants saw the determiner (e.g. '*tib*'), compared to the suffixes (e.g. *'eem'* and *'esh'*) in the input. As in Experiments 1 and 2 *one* determiner was seen across eight stems, but *one* suffix was seen across four stems, because of the nature of our stimuli the more reliable (i.e. more systematic) determiners were also the more available (i.e. more frequently appearing in the input) mapping. As we saw extraction of the determiner-gender regularity, but not of the suffix-gender regularity, this raises the question of whether this was influenced purely by the greater systematicity (i.e. greater reliability) of the determiner mapping, or because of other factors in the language input (which will be discussed further in Chapter 5).

The aim of Experiments 3 and 4 is to investigate this by making suffixes more systematic than determiners (i.e. the opposite of Experiments 1 and 2). Making the suffixes more systematic, may potentially ease the participants' task of suffix-gender regularity extraction. Regardless of sleep, a more systematic mapping due to its higher strength as a cue should be more likely to be extracted than a less systematic mapping, within the Competition model.

Indeed, using both experiments with humans, and simulations, Kempe and MacWhinney (1998) have demonstrated that cases in German, in comparison to Russian cases, are acquired more slowly, and with more errors, despite being simpler. Noun case-marking aids the understanding of the meaning of a sentence (i.e. which of 2 nouns in a sentence is the agent). As the morphological cues for noun case-marking (nominative/accusative) appear less frequently in the input in German, they are more difficult to acquire. Hence, the more frequently-appearing (i.e. more available) cues are easier to acquire. This is because increased frequency speeds up the strengthening of the case-marking-agent connections. Hence, in our Experiments, making the suffixes more systematic and therefore more frequent in the input in comparison to determiners, given the training language we are using, may ease their acquisition.

Kempe and MacWhinney (1998) also found that in Russian the more reliable case-marking blocked the need for other cues to agenthood, while in German the less reliable case-marking resulted in increased reliance on other cues. Hence, in our Experiments 1 and 2, the more systematic determiners may have blocked the acquisition of the less systematic suffixes. Swapping this systematicity may mean that the relatively more systematic suffixes will now be extracted, but block the extraction of the less systematic determiners.

This is a particularly interesting point to explore further, given that the lack of suffix-gender regularity extraction in our Experiments 1 and 2, was in contrast with the existing evidence suggesting that both infants and adults are able to extract statistical regularities from phonological sequences, even without semantic cues (e.g. Gomez et al., 2006; Saffran, Aslin & Newport, 1996; St Clair, Monaghan & Ramscar, 2009). Additionally, van den Boss, Christiansen and Misyak (2012) have shown that a semantic referent which is always consistent with the non-adjacent dependency boosts acquisition beyond just phonology alone.

The extraction of the fully systematic determiners, particularly in the presence of the consistent semantic referents (i.e. the pictured characters) in Experiments 1 and 2, was consistent with the above evidence. The lack of full systematicity in suffixes may have made suffix acquisition more difficult in Experiments 1 and 2. As a reminder, there were 2 determiners and 4 suffixes, hence the determiner "*tib*" always occurred with female pictures. Hence, the determiner-gender co-occurrence was always fully systematic. Yet, the co-occurrence of determiners to suffixes was not fully systematic (determiner "*tib*" could be seen with either suffix "*esh*" or suffix "*eem*"). The co-occurrence of the natural gender to suffixes was also not fully systematic (feminine pictures could be seen with either suffix "*esh*" or suffix "*eem*"). The full systematicity of the determiner-gender regularity may have made that mapping easier to acquire, while the lack of full systematicity within suffix-gender regularities may have made suffix acquisition more difficult. Indeed, Van den Bos and Poletiek (2015) (using a more complex grammar than in our studies) have shown that while semantic cues aid acquisition beyond phonology alone, they are not enough to compensate for increased difficulty of a complex grammar.

Sleep-related consolidation and systematicity

As already discussed in the previous Chapters, the idea that the more systematic mapping may not need additional sleep-related consolidation to boost extraction can be explained in terms of the CLS (e.g. McClelland et al., 1995) and the two-stage systems consolidation models (e.g. O'Reilly & Norman, 2002; Rasch & Born, 2007). On the basis of such models, it can be predicted that the additional benefit of sleep over the equivalent time awake is in sleep's mediation of the hippocampal-neocortex information dialogue, as a result of which initially new information is permanently integrated with existing schemas.

As argued in earlier Chapters, the systematic mappings may be less hippocampally-dependent, and therefore less reliant on this hippocampal-neocortex dialogue. Specifically, if regularities have been fully integrated in the neocortex during the exposure and/ or wake delay, there is no need for additional sleep-associated consolidation (e.g. O'Reilly & Norman, 2002; Rasch & Born, 2007). Therefore, the more systematic mappings may be in turn less reliant on sleep.

4.2 Experiment 3

Aims and predictions

To summarize our aims for Experiments 3: firstly, it is of interest how the combination of factors such as cue reliability (i.e. systematicity) and cue availability (i.e. frequency) within language mappings is related to the extraction of regularities within those mappings. Secondly, it is of interest whether the role of sleep in extraction of regularities is influenced by this combination of factors.

Experiment 3 used the same design and tasks as Experiments 1 and 2. The primary modification was that the systematicity (i.e. cue reliability) and hence frequency (i.e. cue availability) of determiners to suffixes was swapped. In Experiments 1 and 2 there was 1 determiner and 2 suffixes *per* gender. In Experiment 3 this was changed to 2 determiners and 1 suffix *per* gender. The determiners "*tib*" and "*jov*" and the suffix "*eem*" mapped onto feminine items. The determiners "*paz*" and "*ked*" and the suffix "*ool*" mapped on to masculine items (e.g. "*tib rutcheem*"-"bride"; "*jov mofeem*"-"waitress", "*paz soidool*"-"footballer", "*ked cassool*"-"doctor"). This made the suffix mapping cue both more systematic (i.e. reliable) and frequent (i.e. available) within the input. This combination of factors can potentially increase the possibility of extraction. Moreover, the application of this design to Experiment 3 means that we can explore the extent to which a combination of these factors may facilitate suffix-gender regularity extraction.

To summarize, in Experiment 3 we are addressing the issue of suffix-gender regularity extraction, on the one hand. We are doing this by increasing frequency (i.e. cue availability), to make suffix frequency in Experiment 3 comparable to determiner frequency in Experiments 1 and 2. On the other hand, we are still testing the predictions of the CLS regarding systematicity and thus predicting the effect of sleep for the arbitrary but not for systematic mappings.

4.2.1 Method

Participants

The participant criteria were the same as in Experiments 1 and 2. There were a total of 50 participants in Experiment 3. Out of these, twenty-five participants with a mean age of 19.44 (± 1.55 [SD]) were in the sleep group (14 male, 2 left handed). Twenty-five participants with a mean age of 19.30 (± 2.37 [SD] were in the wake group (14 male, 2 left handed). There were 12 participants in the sleep group and 11 participants in the wake group who played either a string instrument or the piano (according to answers on session 1 questionnaire). In return for participation, £12 or 2 hours undergraduate psychology course credit was given.

Design

The design was the same as in Experiment 1. A daytime-wake and an overnight-sleep (at home) group was included (between-subjects design). The sleep group began session 1 at 8.30 PM. The wake group began session 1 at 8.30 AM. Session 2 occurred 12 hours after session 1, such that only participants from the sleep group slept in the delay between session 1 and session 2.

Materials

The materials were the same as in Experiments 1 and 2, with one key exception: there were 4 determiners and 2 suffixes (and in Experiments 1 and 2 there were 2 determiners and 4 suffixes). The determiners *'tib'* and *jov'* and the suffix *'eem'* mapped onto one gender. The determiners *'ked'* and *'paz'* and the suffix *'ool'* mapped onto a different gender. The previously unused determiners *'jov'* and *'paz'* were selected from the ARC nonword database (Rastle, Harrington & Coltheart, 2002). The items were created by using the same stems as in Experiments 1 and 2 and then adding the determiners and suffixes accordingly. As with Experiments 1 and 2, overall there were 16 different words, 8 per gender. As a reminder, in Experiments 1 and 2, there were four stems per suffix, but eight stems per determiner. This was the other way round in Experiment 3. In other words, there were now four stems per determiner, but eight stems per suffix. Spoken versions of all items were recorded by a male native English speaker.

Training items

As with Experiments 1 and 2, participants learnt one of two language versions which differed in the phonological form-gender mappings. In language version one, the determiners *'tib'* and *'jov'* and the suffix *'eem'* mapped on to male gender, the determiners *'ked'* and *'paz'* and the suffix *'ool'* onto female gender, vice versa in language two (determiners *'tib', 'jov',* suffix *'eem'* = female, determiners *'ked', 'paz',* suffix *'ool'* = male).

As with Experiments 1 and 2, there were altogether 192 items in the training task (i.e. the word-picture matching task). This was because each of the 16 words was seen 8 times with its target picture. Each of the 16 words was also seen with 4 different noise pictures (2 of which were of the target gender, the other 2 were of the opposite gender). This is the same ratio of target to noise items as with Experiments 1 and 2.

Generalization items

As with Experiments 1 and 2, two sets of generalization items were used. One set of generalization items was used for the suffix generalization task, which tested participants' extraction of the suffix-gender regularities. The second set of generalization items was used for the determiner and suffix generalization task, which tested participants' extraction of the determiner-gender regularities. As a reminder, the two tasks exposed participants to new items (i.e. untrained word-picture mappings), and tested participants' endorsement of items which were consistent and inconsistent with the trained regularities.

As an example, during training participants learned that the determiner *'tib'* and the suffix *'eem'* map on to female gender pictures. The determiner *'ked'* and the suffix *'ool'* map onto male gender pictures. In that case, in the suffix generalisation task, an example of a consistent (with trained regularities) item would be "*tib jurbeem"* – "maid", an example of an inconsistent item would be *"ked lertheem"* – "firefighter". Note that in the suffix generalisation task the determiner was always consistent with the gender of the pictured character (as with Experiments 1 and 2). In the determiner and suffix generalisation task, an example of a consistent item would be "*tib zimbeem"* – "cheerleader", an example of an inconsistent item would be "*ked felchool"* – "nun".

As with Experiments 1 and 2, there were 16 items per task, half masculine, half feminine. The same stems were used as in the two generalization sets from Experiments 1 and 2, and then the determiners and suffixes were added on accordingly. As with Experiments 1 and 2, half of the items in each set were consistent with the trained regularities, and the other half were not.

Procedure

The participants were tested on the same tasks as in Experiment 1, presented in the same order. The only difference was in the determiner selection task, as there were now 4 determiners, instead of 2. Thus, the response choices in this task were changed to the four determiners. Specifically, participants were asked to indicate which of the four choices of determiners matched the word and the picture. Participants used the 4 buttons on the buttonbox (buttons X, Y, B and A, see Figure 24) to respond.

Figure 24. Buttonbox layout.

The arrangement of determiners to buttons was counterbalanced across participants. ‘KED’, ‘PAZ’, ‘JOV’ and ‘TIB’ remained at the bottom of the screen (beneath the picture and the word) during each trial to remind participants which button corresponded to which response (see Figure 25 for examples).

	KED			TIB	
JOV		TIB	PAZ		KED
	PAZ			JOV	

Figure 25. Examples of determiner-button arrangement on the determiner selection task.

4.2.2 Results

Language tasks

Training task (Session 1)

As with Experiments 1 and 2, participants were trained on the artificial language using a word-picture matching task. Also in line with our analyses of Experiments 1 and 2, both the accuracy and the RT analyses will be presented to provide an indication of learning of the language.

Accuracy analysis

Participants’ accuracy at the end of training was measured to assess their knowledge of the mappings and compared between the experimental groups and the two language versions. As this was the participants’ first exposure to the language, it was expected that there would be no significant differences between groups at the end of training.

A two way between subjects ANOVA was run with the independent variables of group (sleep or wake) and language (version one or two). The dependent variable was accuracy (proportion of correct responses for the “target” trials only) at the final block of the training task. There was no significant main effect of group ($F(1, 46) = 0.02$, $p = .895$) or of language version ($F(1, 46) = 2.60$, $p = .113$) and no significant interaction between group and language ($F(1, 46) = .05$, $p = .825$). As shown in Figure 26, participants from both groups (sleep and wake) and both language versions gained a similar level of knowledge of the language by the end of training.

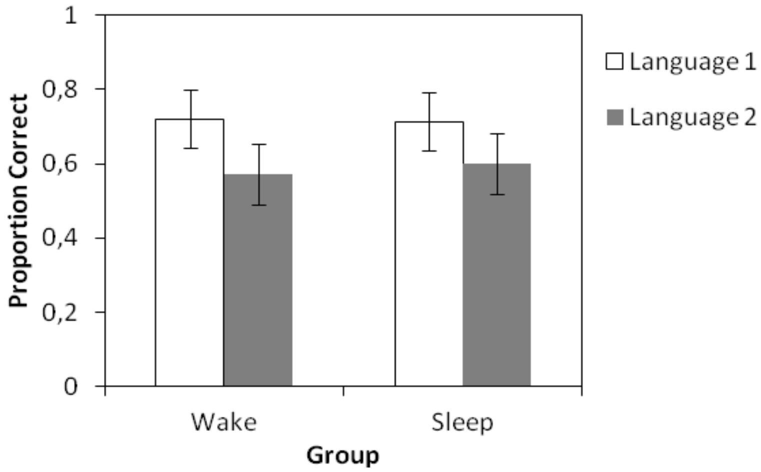


Figure 26. Experiment 3 performance on the last block of the training task for the two groups and the two versions of the training language.

Correct RTs analysis on the training task

The same analyses of RTs as in Experiments 1 and 2 were run. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake and sleep). First, the dependent variable was the change in target RTs compared to block 1.There was a significant main effect of block ($F(1, 47) = 4.62$, $p = .012$), no significant main effect of group ($F(1, 47) = .00$, $p = .981$), and no significant interaction between block and group ($F(1, 47) = 1.12$, $p = .332$).

In addition, there was no significant main effect of language version ($F(1, 47) = .00$, $p = .993$) and no significant interaction between group and language ($F(1, 47) = .16$, $p = .693$). There was also no significant interaction between block and group ($F(1, 47) = 1.12$, $p = .332$), a marginally significant interaction between block and language ($F(1, 47) = 2.51$, $p = .087$), and no significant interaction between block, group and language ($F(1, 47) = .31$, $p = .736$).

As shown in Figure 27, all participants got faster with training, regardless of group, as expected. Importantly, this suggests that the initial level of acquisition was similar for all participants.

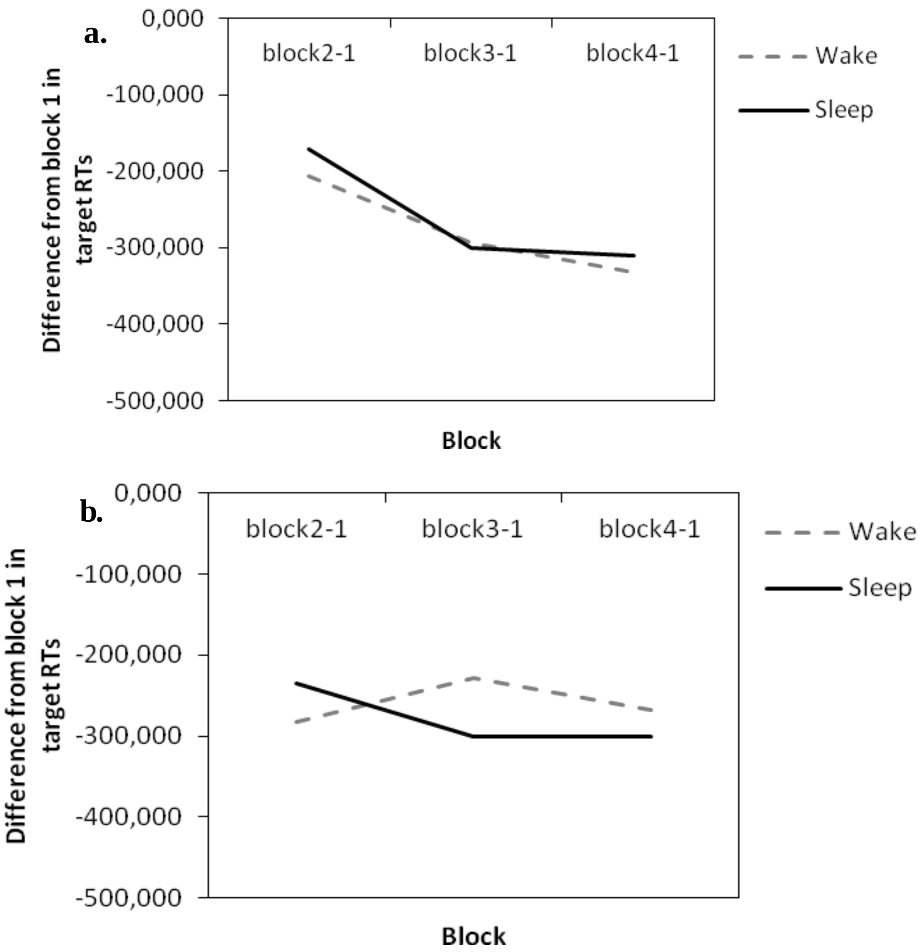
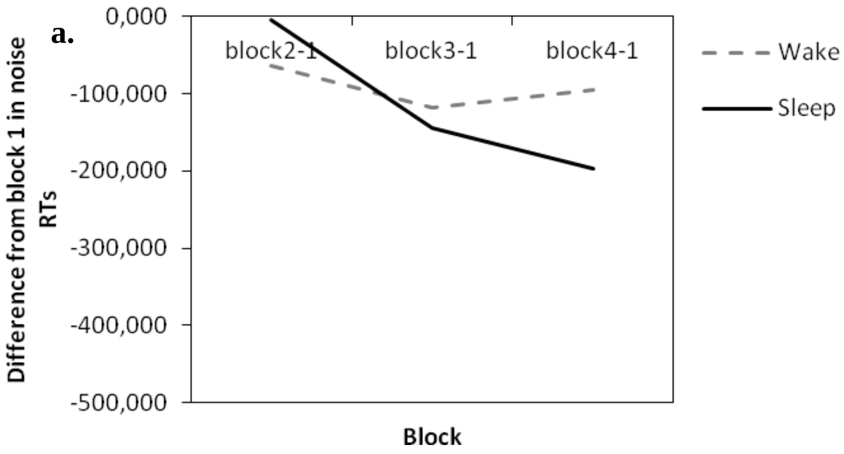


Figure 27. Difference in target RTs from block 1 (Figure 27a. shows language version 1 RTs, Figure 27b. shows language version 2 RTs).

Secondly, the dependent variable was the change in noise RTs compared to block 1. There was a significant main effect of block ($F(1, 47) = 4.37$, $p = .015$), no significant main effect of group ($F(1, 47) = .01$, $p = .923$), and no significant interaction between block and group ($F(1, 47) = 1.77$, $p = .176$).

In addition, there was no significant main effect of language version ($F(1, 47) = .98$, $p = .326$), and no significant interaction between group and language ($F(1, 47) = .02$, $p = .884$). There were also no significant interactions between block and group ($F(1, 47) = 1.77$, $p = .176$), block and language ($F(1, 47) = .98$, $p = .381$), or block, group and language ($F(1, 47) = .64$, $p = .530$). As shown in Figure 28, all participants got faster with training, and the difference in RTs compared to block 1 increased with training, regardless of group.



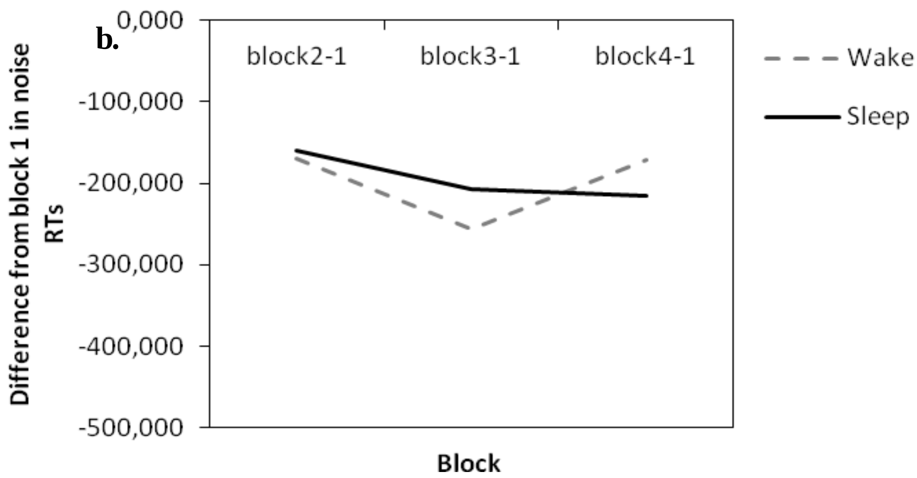


Figure 28. Difference in noise RTs from block 1 (Figure 28a. shows language version 1 RTs, Figure 28b. shows language version 2 RTs).

Overall, the results from the accuracy analysis from the training task suggested that any differences found between the groups in session two language tests cannot be due to differences in initial acquisition levels. As both language versions had a similar level of performance, all other tasks will be analysed collapsing across language version. The results from the RT analysis also suggested that the initial level of acquisition was similar across all participants, regardless of group.

Language tests (Session 2)

Most arbitrary mapping: vocabulary

Translation Recognition Task

As with Experiments 1 and 2, participants' memory for the trained words and their meanings was tested using a translation recognition task. The mean accuracy in the sleep group was 80.5% for match trials, and 81.3% for mismatch trials. The mean accuracy in the wake group was 80.0% for match trials, and 83.3% for mismatch trials. This suggests that participants in both groups reached a similarly good level of performance.

A between subjects t-test was run with the independent variable of group (sleep or wake) and A' discrimination as the dependent variable. There were no significant differences between groups ($t(48) = -.06$, $p = .950$). As shown in Figure 29, all participants had good discrimination between match and mismatch trials, regardless of group. There was no evidence for a benefit of sleep on this task, consistent with Experiments 1 and 2.

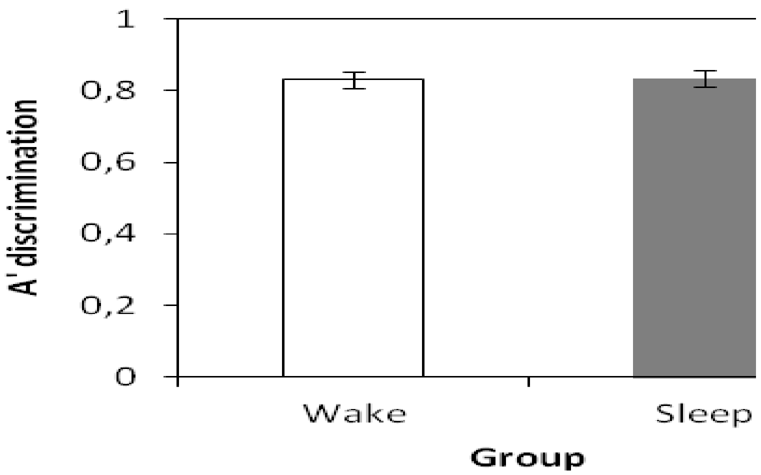


Figure 29. Experiment 3 performance on the translation recognition task.

Recall (stems)

As with Experiments 1 and 2, participants' memory for the trained word-picture mappings was also tested using a picture naming recall task. Each of the trained novel words was cued twice: once with the picture only, the second time with the initial grapheme of the word (e.g. *'b'* for *'biseem'*). Crucially, the focus of the analysis on this task is on any emerging group differences, as opposed to differences as a result of cue type. It was expected that there would be a benefit of sleep for stem recall, consistent with Experiments 1 and 2, and the predictions of the CLS model (e.g. McClleland et al., 1995).

The stems only and separately the suffixes only were extracted from the transcriptions of participants' responses to be analysed separately. This was done in order to separate each word into its entirely arbitrary component (the stem) and its systematic (by design) component (the suffix), in the same way as in Experiments 1 and 2.

To analyse recall of the arbitrary element of the words, i.e. the stems, a mixed ANOVA was run with the between subjects variable of group (sleep or wake) and the within subjects variable of cue type (picture and grapheme, or picture only). The dependent variable was stem accuracy (proportion correct out of total productions). Participants from both groups recalled a similarly small proportion of the novel word stems (Figure 30).

There was no significant main effect of group ($F(1, 48) = 2.24$, $p = .141$). There was a significant main effect of cue type ($F(1, 48) = 13.80$, $p < .002$), and no interaction between group and cue type ($F(1, 48) = 1.63$, $p = .208$). As Figure 30 shows, participants from both groups produced more stems correctly when they were cued with both a picture and a grapheme than just the picture only, regardless of sleep. Although numerically participants from the sleep group had higher performance than participants from the wake group, this difference was not significant.

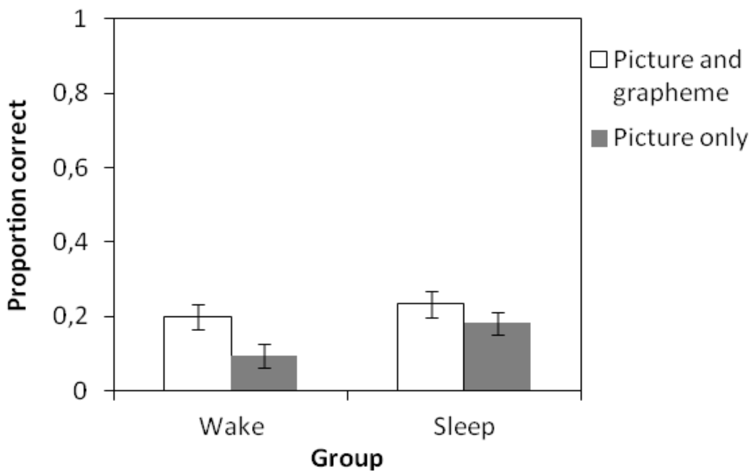


Figure 30. Experiment 3 stem accuracy on the recall task.

Overall, now that the systematicity of the mappings has been swapped, the analysis of the arbitrary mappings (the translation task and stem recall task) has not provided any evidence of a benefit of sleep for consolidation (as there was no significant main effects of group neither in the translation, nor in the recall task), unlike Experiments 1 and 2.

Most systematic mapping: suffixes (grammar)

Recall (suffixes)

As with Experiments 1 and 2, to assess participants' memory of the systematic element of the word, the recall accuracy of the suffixes (which were separated from stems in the transcriptions) was analysed. It was expected that there would be a benefit of sleep for suffix recall, consistent with Experiments 1 and 2. Overall participants in both groups recalled more suffixes correctly than the stems, as seen in Figure 31 and Figure 30. This is in line with Experiment 1 performance (i.e. prior to the swapping the systematicity).

A mixed ANOVA with a between subjects factor of group and within subjects factor of cue type was run with suffix accuracy (proportion correct out of total productions) as the dependent variable. There was a significant main effect of group ($F(1, 48) = 6.86$, $p = .012$), and no significant main effect of cue type ($F(1, 48) = .581$, $p =.450$). There was no significant interaction between group and cue type ($F(1, 48) = .138$, $p = .711$).

As Figure 31 shows, participants in the sleep group produced more suffixes correctly than participants in the wake group. Although suffixes are a systematic mapping in the input, participants who have not extracted this systematicity, may be treating suffixes as individual mappings. In other words, participants may be learning each picture-suffix mapping individually, as opposed to extracting the suffix-gender regularity mapping. Hence, if suffixes are treated as an individual, arbitrary mapping within the recall task, the benefit of sleep over wake in this task reflects the benefit of sleep for the consolidation of individual suffixes at the arbitrary level. Analysis of the suffix generalisation task will provide evidence whether the higher suffix recall is driven by better consolidation of individual arbitrary suffix-picture mappings, or indicates extraction of the suffix-gender regularity.

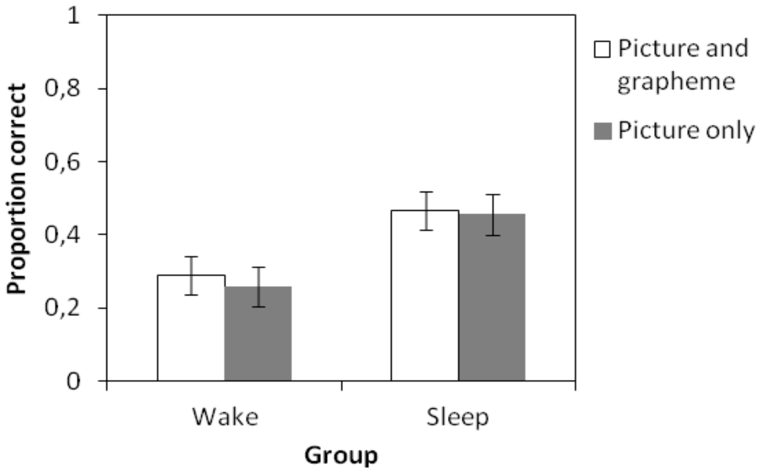


Figure 31. Experiment 3 suffix accuracy on the recall task.

Suffix generalisation task

To test if participants extracted the suffix-gender mappings which they were exposed to during training, a suffix generalisation task was used. This task was the same as in Experiments 1 and 2. Participants' endorsement (i.e. pressing 'match') for consistent and inconsistent items was assessed. Crucially, it is the endorsement pattern that is of interest, as opposed to overall performance. A pattern whereby consistent items are endorsed significantly more than inconsistent items would reflect extraction. In line with Experiments 1 and 2, and predictions of the CLS (e.g. McClleland et al., 1995), we are not expecting to see strong evidence of a benefit of sleep for extraction. In addition, now that the determiners have been made less systematic, while the suffixes have been made more systematic in relation to Experiments 1 and 2, it was of interest whether we would now observe evidence of suffix-gender regularity extraction.

A mixed ANOVA was run with the between subjects variable of group (sleep or wake) and within subjects variable of item consistency (mappings consistent or inconsistent with trained regularities). Proportion endorsed (pressing 'match' to consistent and inconsistent items) was the dependent variable. There was a significant main effect of consistency ($F(1, 48) = 7.00$, $p = .011$), no significant main effect of group ($F(1, 48) = .742$, $p = .393$), and no significant interaction between group and consistency ($F(1, 48) =.691$, $p = .410$). Crucially, inconsistent items had higher endorsement rates than consistent items (as shown in Figure 32). This pattern was not in line with what would be expected if the suffixes had been extracted.

One sample t-tests against chance may provide further clarification to this result. Unlike in Experiments 1 and 2, these findings suggest that inconsistent items were more likely to be endorsed than consistent items. However, in both types of items, the endorsement rate was low, and close to the chance level of 0.5.

One sample t-tests against chance (0.5) revealed that participants in the sleep group were at chance for endorsing both consistent and inconsistent items (consistent (t (26) = -.362, p =.721), inconsistent (t(24) = 1.71, p = .100)). Participants in the wake group endorsed consistent items at chance but endorsed inconsistent items above chance (consistent (t (24) = 1.20, p = .241), inconsistent (t(24) = 2.77, p =.011)). As Figure 32 shows, participants from both groups were more likely to endorse inconsistent items than consistent. The findings about suffix-gender regularity extraction are inconclusive at this stage.

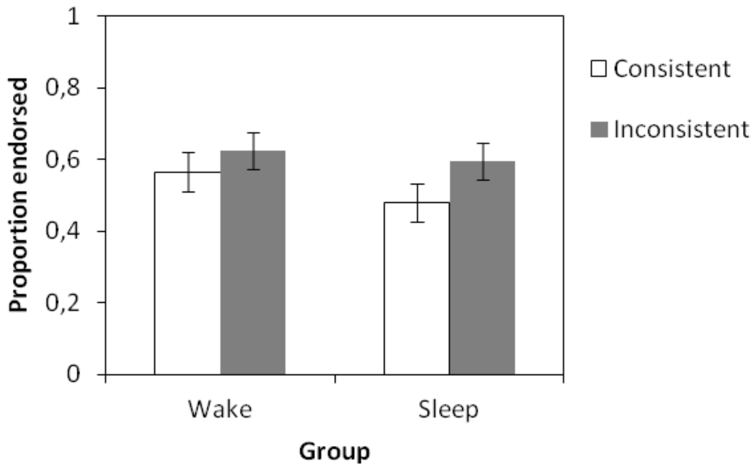


Figure 32. Experiment 3 performance on the suffix generalisation task.

To summarize, the findings from the suffix recall task suggest that a 12 hour delay between training and testing which included a period of sleep was beneficial for consolidating the individual picture-suffix mappings. The findings from the suffix generalisation task were inconclusive, and did not provide enough evidence to indicate suffix-gender regularity extraction, despite the suffixes now being fully systematic. The pattern of endorsement was in the opposite direction from what would be expected if the suffix-gender regularity was extracted.

Intermediate systematicity mapping: determiners (grammar)

Determiner selection task

As with Experiments 1 and 2, participants' memory for the trained determiner mappings was assessed using a determiner selection task. In order to select the correct determiner out of four choices in this task, participants could rely on their familiarity of the individual phonological forms of the determiner-novel word combination (e.g. that the determiner 'ked' should co-occur with the novel word 'cassool'). Alternatively, participants could rely on the individual 'determiner-novel word' to picture mapping (e.g. that 'ked cassool' should be seen with a picture of a doctor). Alternatively, participants could have extracted the determiner-gender regularity (e.g. that the determiner 'ked' should co-occur *only* with male gender pictures). Participants' accuracy in selecting which one of the four trained determiners matched the presented word-picture pairing was assessed.

A between subjects t-test was run with the independent variable of group (sleep or wake) and the dependent variable of accuracy (proportion correct). There was no significant difference between groups ($t(48) = .139$, $p = .890$). One sample t-tests against chance (0.25, given 4 determiners) revealed that participants in both groups were above chance at correctly selecting determiners (wake ($t(24) = 4.25$, $p < .001$), sleep ($t(24) = 3.69$, $p =< .001$)). Despite being above chance, memory for individual determiner-picture mappings was low for both groups of participants. This is reflected by the low absolute performance. More specifically, the mean number of correctly selected determiners out of 16, was 6.46 for the wake group, and 6.23 for the sleep group, i.e. just 38-40%.

As Figure 33 shows, participants in both groups reached a similarly low level of performance. Nonetheless, participants in both groups learned the individual determiner mappings above chance, indicating that memory traces for the individual determiner-picture mappings should be available. There was no evidence for an additional benefit of sleep for individual determiner-picture mapping consolidation.

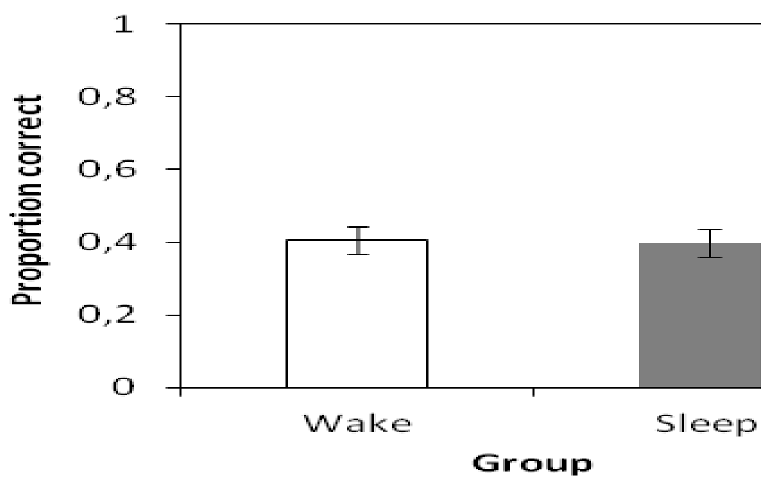


Figure 33. Experiment 3 participants' memory for trained determiner-picture mappings assessed using the determiner selection task.

Analysis of incorrect responses on the determiner selection task

As already mentioned, participants may have relied on either the memory of the *individual* determiner-picture mappings, or on the extracted determiner-gender *regularities* in order to complete the task. To investigate this, an analysis of errors in this determiner selection task was performed. This analysis of errors was similar to the analysis of suffix errors on the suffix recall task in Experiments 1 and 2. Specifically, errors reflecting correct determiner-gender mappings (proportion of correct gender determiners out of the incorrect but trained determiners) were compared against the chance level of 0.33. This is because given that there were four determiners, each response could have three possible incorrect trained determiners and only one of the three could be of the correct gender, given that there were two determiners per gender.

Reliance on the extraction of the determiner to picture gender mapping would be reflected by correct determiner-gender mappings being produced above chance. It was of interest whether now that the determiners were made less systematic than in Experiments 1 and 2, there would be evidence of reliance on the determiner-gender regularity extraction.

As shown in Table 9, all of the participants were at chance at selecting the correct determiner-gender mapping when making an error. This suggests that in addition to the participants' low memory for the individual determiner-picture mappings, there was no evidence of participants relying on the extracted determiner-gender regularity to complete the task.

Table 9. Results of correct determiner-gender errors analysis in the determiner selection task using one sample t-tests.

Group	t	df value	*p-value*	Mean proportion
sleep	-.75	24	.463	.31
wake	-.43	24	.672	.32

The analysis of the determiner selection task has not provided any evidence that sleep is beneficial for the consolidation of the individual determiner-picture mappings, regardless of the level of their systematicity. It has also not provided any evidence that the determiner-gender regularities were extracted, now that the determiners were made less systematic, and the suffixes more systematic. Moreover, individual determiner-picture mapping memory in terms of absolute performance, although above chance, was nonetheless low.

Determiner and suffix generalisation task

To test further if participants extracted the determiner-gender mappings which they were exposed to during training, a determiner and suffix generalisation task was used, as with Experiments 1 and 2. This was also a word-picture matching task like the suffix generalisation task, with novel items. For the inconsistent mappings, the determiner and the suffix were inconsistent with the natural gender of the depicted character relative to the trained mappings (e.g. consistent with trained: "*jov vedeem*" – "fortuneteller", inconsistent: "*jov gatcheem*" – "fencer". Participants' endorsement (i.e. pressing 'match') for consistent and inconsistent items was assessed. As a reminder, the pattern of endorsement was of interest here as the indication of generalisation.

A mixed ANOVA was run with the between subjects variable of group (sleep or wake) and within subjects variable of consistency (consistent or inconsistent). The dependent variable was the proportion endorsed. There was no significant main effect of consistency ($F(1, 48) = .04$, $p = .838$), no significant main effect of group ($F(1, 48) = 1.01$, $p = .321$), and no significant interaction between group and consistency ($F(1, 48) = .23$, $p = .633$). As shown in Figure 34, there was no evidence of determiner-gender regularity extraction, and no evidence of an additional benefit of sleep.

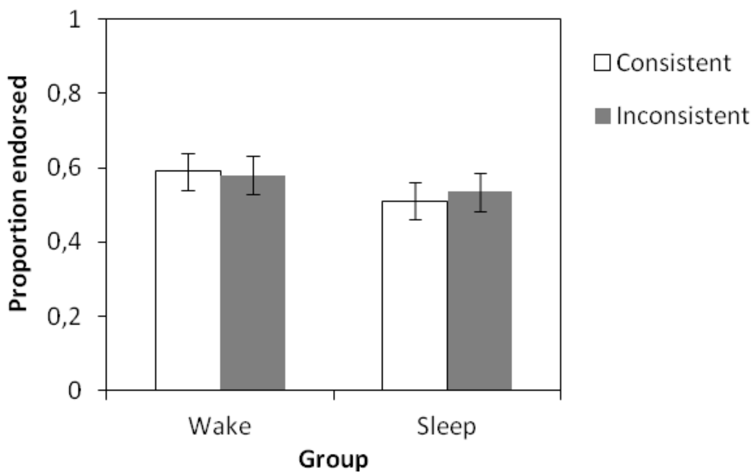


Figure 34. Experiment 3 participants' ability to generalise determiner-gender regularities assessed using the determiner and suffix generalisation task.

One sample t-tests against chance (0.5) revealed that participants in both groups were at chance at endorsing both consistent items (wake: ($t(24) = 1.67$, $p = .107$), sleep: ($t(24) = .22$, $p = .826$)), and inconsistent items (wake: ($t(24) = 1.54$, $p = .137$), sleep: ($t(24) = .66$, $p = .514$)). This suggests that participants in both groups were not influenced by the consistency of the trained regularities when endorsing new picture-word pairs, unlike in Experiments 1 and 2.

Moreover, it may be important to further highlight that the performance was close to chance. Even if the participants had not extracted the determiner-gender regularity, participants have already been exposed to the determiners and suffixes. Hence, they could have relied on purely their *phonological* (as opposed to semantic mapping) memory for the determiners and suffixes. To elaborate, if participants had a good memory of the trained determiners and suffixes *phonologically,* they would have had a higher *overall* rate of performance on this task, regardless of the endorsement *pattern.* Given that performance was close to chance, this further supports the findings from the determiner selection task, that participants' memory for the individual determiner-picture mappings was low.

To summarize the findings from the determiner tests, now that the determiners were less systematic, while suffixes were more systematic, there was evidence of the memory for individual determiner-picture mappings being low, although above chance. There was no evidence of the extraction of the determiner-gender regularities. Despite the low individual memory, consistently with Experiment 1 (when memory was higher), there was still no additional benefit of sleep for the consolidation and extraction of the determiners. This suggests that neither the level of performance, nor the systematicity of the mapping affected sleep-dependent consolidation.

Comparison of Experiments 1 and 3

The key difference between Experiments 1 and 3 was the language type. In Experiment 1, the language type was characterised by more systematic determiners and less systematic suffixes. In Experiment 3, the language type was characterised by less systematic determiners and more systematic suffixes. To explore the differences in learning for the two types of languages, performance in Experiment 3 was compared with performance in Experiment 1. This variable was labelled language type, given that this was the key differences between the two Experiments. Given that there were no systematic differences between the two versions of the language, language data will be collapsed across the two versions.

Training task (session one)

In order to investigate whether decreasing the level of systematicity of the determiners and increasing the level of systematicity of the suffixes affected the acquisition of the language mappings, performance between Experiments 1 and 3 was compared. Performance at training was compared firstly in terms of accuracy, and secondly in terms of correct RTs for target and noise trials separately. These were the same analyses that were previously carried out on Experiment 1 and Experiment 3 data separately. These analyses will collapse across the two language versions.

Accuracy analysis

To analyse accuracy at the end of training, a two way between-subjects ANOVA was run. The between subjects variables were group (wake/ sleep) and language type (Experiment 1/ Experiment 3). The dependent variable was accuracy (proportion correct) at the end of the training task. There was no significant main effect of Experiment (F(1, 100) = 1.12, p = .296), no significant main effect of group (F(1, 100) = .13, p = .722), and no significant interaction between Experiment and group (F(1, 100) = .29, p = .589). As shown in Figure 35, participants in both Experiments and in both groups had a similar level of performance at the end of training.

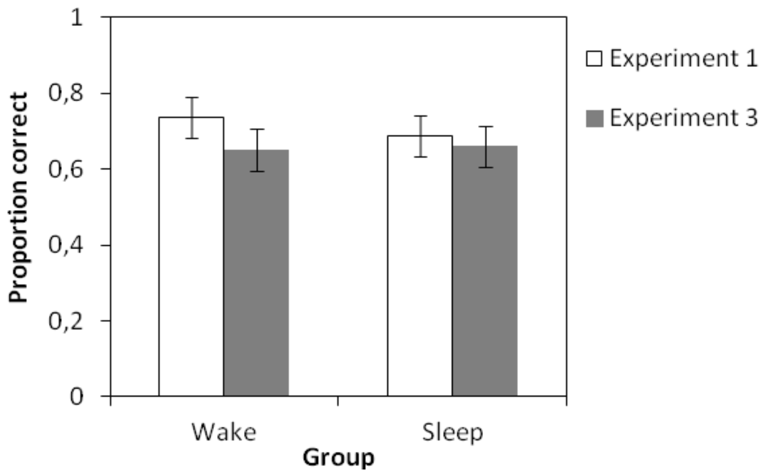


Figure 35. Performance on the training task for Experiments 1 and 3.

Correct RTs analysis on the training task

The RT data for correct responses was analysed separately for target trials and separately for noise trials. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), and the between subjects variables of Experiment (Experiment 1/ Experiment 3) and group (sleep/wake). First, the dependent variable was the change in target RTs compared to block 1.

There was no significant main effect of Experiment ($F(1, 100) = 1.23$, $p = .271$), and no significant main effect of group ($F(1, 100) = .05$, $p = .822$). There was a significant main effect of block ($F(1, 100) = 19.27$, $p < .001$), and no significant Experiment x block interaction ($F(1, 100) = .57$, $p = .568$). There was also no significant interaction between block and group ($F(1, 100) = 1.76$, $p = .174$), no significant interaction between group x Experiment ($F(1, 100) = .05$, $p = .823$). There was also no significant interaction between block x Experiment x group ($F(1, 100) = .57$, $p = .564$).

This suggests that participants from both Experiments, both from the wake groups (Figure 36a) and the sleep groups (Figure 36b) improved a similar amount across the blocks, indicative of similar levels of learning, as shown in Figure 36.

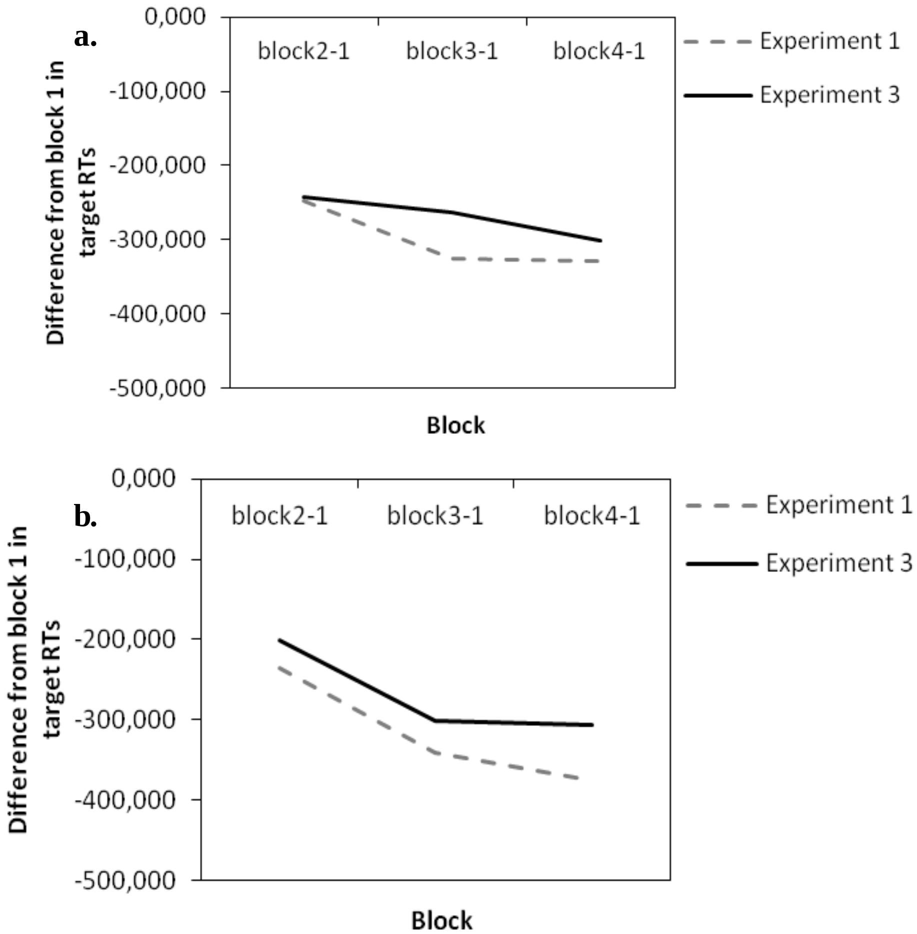


Figure 36. Difference from block 1 in target RTs in Experiments 1 and 3 (Figure 36a. shows wake group's RTs, Figure 36b. shows sleep group's RTs).

Next, the RT data for correct responses was analysed separately for noise trials. As mentioned in Chapter 2, 1 participant from Experiment 1 was not included in the analysis, as they did not have any correct responses for noise RTs in some of the blocks. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1), and the between subjects variables of and Experiment (Experiment 1/ Experiment 3). The dependent variable was the change in noise RTs compared to block 1.

There was a marginal main effect of Experiment ($F(1, 99) = 3.02$, $p = .085$), no significant main effect of group ($F(1, 99) = .21$, $p = .648$), a significant main effect of block ($F(1, 99) = 13.26$, $p < .001$), and no significant Experiment x block interaction ($F(1, 99) = 1.45$, $p = .238$). There was a significant interaction between block and group ($F(1, 99) = 3.29$, $p = .039$), and no significant interaction between group x Experiment ($F(1, 99)$

= 1.00, *p* = .759). There were also no significant three-way interaction between block x Experiment x group ($F(1, 99) = .09$, $p = .919$).

This suggests that participants from both Experiments improved a similar amount across the blocks, although with marginally greater improvement in Experiment 1, than in Experiment 3, indicative of similar levels of learning, as shown in Figure 37. In addition, participants in the sleep group (Figure 37b) improved more across the blocks than participants from the wake group (Figure 37a), suggesting that learning might have been greater in the sleep groups, although this was not reflected in the target RTs (as mentioned above).

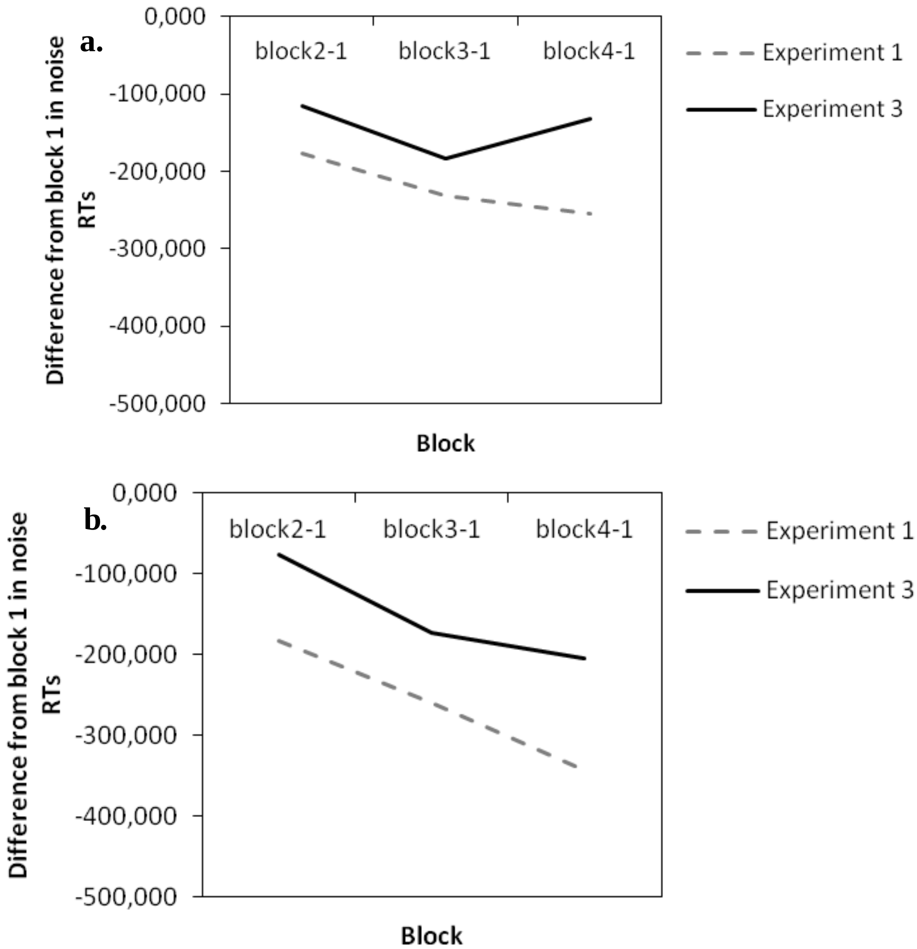


Figure 37. Difference from block 1 in noise RTs in Experiments 1 and 3 (Figure 36a. shows wake group's RTs, Figure 36b. shows sleep group's RTs).

To summarize, based on the accuracy analysis of the training task, there was no evidence to suggest that swapping the systematicity of the determiner and suffix mappings affected initial acquisition. This means that any differences which may be found between language types in session two tasks would be as a result of the impact that swapping systematicity had on the processes of consolidation, as opposed to initial acquisition. The analysis of the correct target RTs suggested that participants from both groups and both Experiments reached a similar level of learning, although analysis of the correct *noise* RTs (but not correct *target* RTs) suggested that participants in the sleep groups (of Experiments 1 and 3) may have overall had greater improvement in RTs across the noise trials than participants in the wake groups (of Experiments 1 and 3).

Language tests (session two)

Most arbitrary mapping: vocabulary

Translation Recognition Task

As a reminder, participants' memory for the trained words and their meanings was tested using a translation recognition task. To analyse if swapping systematicity had any effect on recognition memory, a two way between-subjects ANOVA was run. The between subjects variables were group (wake/ sleep) and language type (Experiment 1/ Experiment 3). The dependent variable was A' discrimination. There was no significant main effect of language type ($F(1, 100) = 1.85$, $p = .177$), no significant main effect of group ($F(1, 100) = .93$, $p = .339$), and no significant interaction between language type and group ($F(1, 100) = .76$, $p = .386$).

As shown in Figure 38, participants in both language types and in both groups had a similarly good level of discrimination between correct and incorrect trials. Hence, there was no evidence to suggest that swapping the systematicity of the determiner and suffix mappings affected recognition memory.

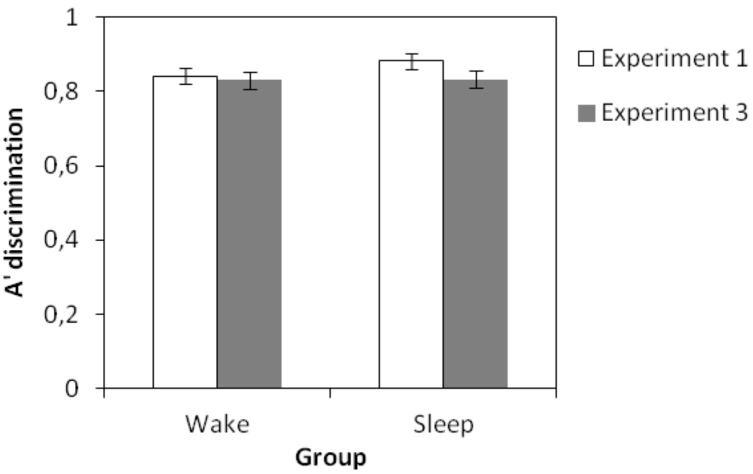


Figure 38. Performance on the translation recognition task for Experiments 1 and 3.

Recall (stems)

As a reminder, participants' memory for the trained word-picture mappings was tested using a picture naming recall task. To investigate whether swapping systematicity impacted stem recall (the individual, arbitrary mapping), a mixed ANOVA was run. The within subjects variable was cue type (picture and grapheme, or picture only). The between subjects variables were group (wake/ sleep) and language type (Experiment 1/ Experiment 3). The dependent variable was proportion of stems recalled correctly (out of total productions).

There was a significant main effect of language type ($F(1, 100) = 35.33$, $p < .001$), and a significant main effect of group ($F(1, 100) = 7.67$, $p = .007$). There was no significant interaction between language type and group ($F(1, 100) = .03$, $p = .853$). As shown in Figure 39, participants in Experiment 3, who were exposed to the less systematic determiners and hence the more systematic suffixes recalled significantly more stems correctly, regardless of group. Potentially, decreasing determiner systematicity and hence increasing suffix systematicity made the stems appear more salient. Moreover, in both language types, participants from the sleep group recalled significantly more stems correctly, than participants from the wake group.

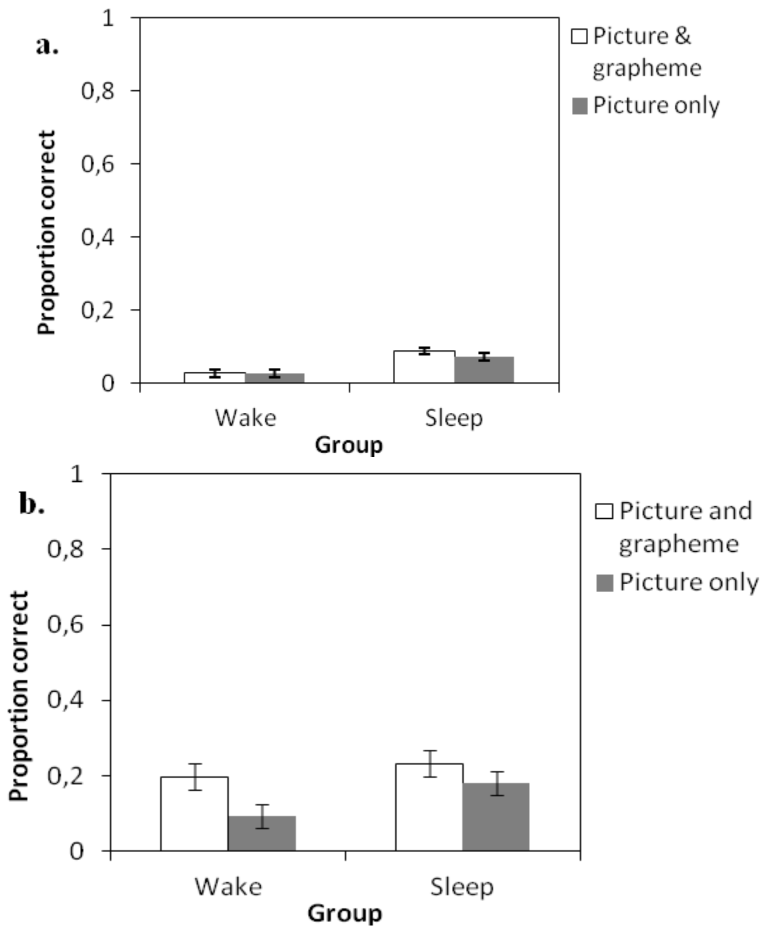


Figure 39. Stem recall accuracy (Figure 39a. shows Experiment 1 accuracy, Figure 39b. shows Experiment 3 accuracy).

Additionally for stem recall, there was a significant main effect of cue type (F(1, 100) = 16.32, p < .001), and a significant interaction between language type and cue type (F(1, 100) = 10.59, p = .002). There were no significant interactions between group and cue type (F(1, 100) = .83, p = .366), or cue type, group and language type (F(1, 100) = 2.55, p = .114).

As shown in Figure 40, the grapheme cue provided more of a boost for the less systematic determiners and more systematic suffixes language type (Experiment 3). In this language type performance was already higher, than for the more systematic determiners and less systematic suffixes language type (Experiment 1). This may be because in Experiment 3, as a result of increasing the suffix systematicity, the suffixes were simplified (i.e. participants had to learn 2 suffixes, instead of 4).

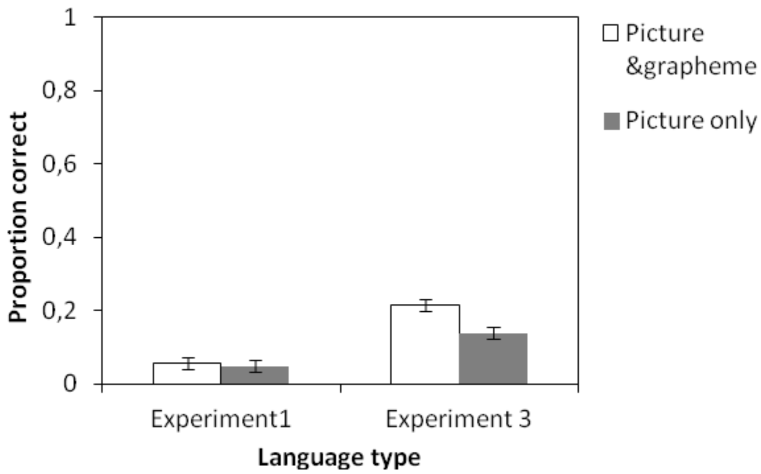


Figure 40. Comparison of stem recall in Experiments 1 and 3 with cue type included.

To summarize the findings comparing performance between the two Experiments (i.e. language types) for the arbitrary mappings, there was no evidence to suggest that language type affected recognition memory. Language type affected stem recall. Stem recall was higher in the language type where determiners were less systematic, but suffixes were more systematic. Regardless of language type, sleep benefited the consolidation of stems in this joint analysis of Experiments 1 and 3. Decreasing the systematicity of the determiners and increasing the systematicity of the suffixes boosted stem recall, potentially by increasing the salience of the stems.

Swapped systematic mapping: suffixes (grammar)

Recall (suffixes)

As a reminder, to assess participants' memory of the systematic element of the word, the recall accuracy of the suffixes (which were separated from stems in the transcriptions) was analysed. To investigate whether swapping systematicity impacted suffix recall, a mixed ANOVA was run. The within subjects variable was cue type (picture and grapheme, or picture only). The between subjects variables were group (wake/ sleep) and language type (Experiment 1/ Experiment 3). The dependent variable was the proportion of suffixes correctly recalled (out of total productions). There was no significant main effect of language type ($F(1, 100) = .16$, $p = .668$). There was a significant main effect of group ($F(1, 100) = 28.85$, $p < .001$). There was no significant interaction between language type and group ($F(1, 100) = 1.72$, $p = .193$).

As shown in Figure 41, participants in both language types overall reached a similar level of suffix recall. Despite having 2 suffixes and hence increased systematicity of suffixes in Experiment 3, as opposed to 4 suffixes and hence decreased systematicity in Experiment 1, statistically recall was still similar across the two Experiments. Overall, participants from the sleep group had higher suffix recall than participants from the wake group, regardless of language type.

Additionally for suffix recall, there was a significant main effect of cue type ($F(1, 100) = 5.89$, $p = .017$), no significant interaction between language type and cue type ($F(1, 100) = 1.04$, $p = .312$). There were also no significant interactions between group and cue type ($F(1, 100) = .03$, $p = .858$), or cue type, group and language type ($F(1, 100) = .75$, $p = .387$). As shown in Figure 41, all participants had higher suffix recall when the grapheme cue was present.

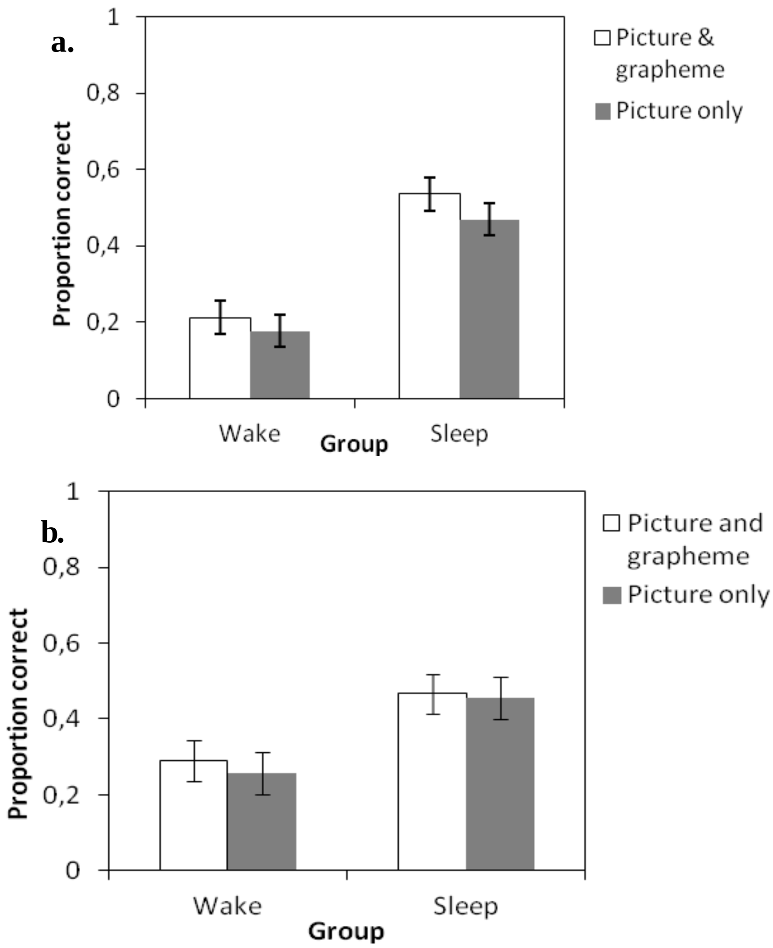


Figure 41. Suffix recall accuracy (Figure 41a. shows Experiment 1 accuracy, Figure 41b. shows Experiment 3 accuracy).

Overall, swapping the systematicity affected stem recall, as stem recall was boosted in Experiment 3, when the determiner systematicity was decreased, and the suffix systematicity was increased. There was no evidence to suggest that swapping the systematicity affected individual suffix recall, despite the increased suffix systematicity.

Suffix generalisation task

As a reminder, the suffix generalisation task was used to assess participants' ability to generalise the trained suffix-gender regularities to previously unseen items. To test if any differences between groups in suffix-gender regularity extraction emerged as a result of differences between the two language types, the performance on the suffix generalisation task was compared between Experiments 1 and 3.

In order to investigate this, a mixed ANOVA was run. The within subjects variable was consistency (items consistent/ items inconsistent with the trained regularities). The between subjects variables were group (wake/ sleep) and language type (Experiment 1/ Experiment 3). The dependent variable was the proportion endorsed (i.e. pressing 'match' to consistent and inconsistent items). Crucially, the pattern of endorsement of consistent versus inconsistent items was of interest here, as indication of generalisation of the trained regularities to new items, as opposed to overall performance.

There was no significant main effect of consistency ($F(1, 100) = 1.82$, $p = .180$). There was no significant main effect of group ($F(1, 100) = .49$, $p = .484$), and no significant interaction between group and consistency ($F(1, 100) = 1.54$, $p = .217$). There was a significant main effect of language type ($F(1, 100) = 4.59$, $p = .035$), and a significant interaction between language type and group ($F(1, 100) = 3.96$, $p = .049$). There was also a significant interaction between language type and consistency ($F(1, 100) = 8.64$, $p = .004$). There was no significant three-way interaction between consistency, group and language type ($F(1, 100) = .01$, $p = .915$).

The key findings from this task relate to the pattern of endorsement (i.e. involve consistency), as opposed to the overall level of performance, regardless of consistency. As shown in Figure 42, and as reflected by the significant language type x consistency interaction, the pattern of endorsement was more in line with suffix-gender regularity extraction in Experiment 1, than in Experiment 3. In Experiment 1, consistent items were endorsed more than inconsistent. In Experiment 3, the pattern of extraction was the opposite of the pattern expected for suffix-gender regularity extraction, as inconsistent items were endorsed more than consistent. There was no evidence that the extraction of suffix-gender regularities differed as a result of group (as there were no significant group x consistency, or consistency x group x language type interactions).

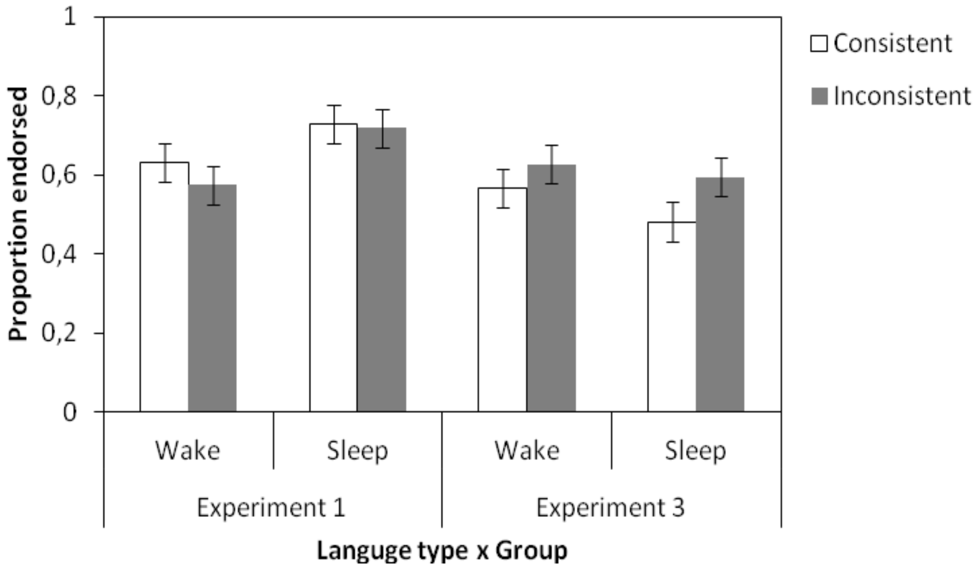


Figure 42. Performance on the suffix generalisation task for Experiments 1 and 3.

To summarize the findings from the suffix tests, there was no evidence to suggest that language type affected individual suffix consolidation (as shown in the suffix recall task). In both Experiments there was a sleep benefit for the recall of individual picture-suffix mappings. The significant language type x consistency interaction on the suffix generalization task is likely due to the lack of differences in endorsement of consistent and inconsistent items in Experiment 1, but greater endorsement of inconsistent compared to consistent items in Experiment 3. Both of these patterns of endorsement indicate no extraction of the suffix-gender regularity.

Swapped systematic mapping: determiners (grammar)

Determiner selection task

As a reminder, participants' memory for the trained determiner-picture mappings was assessed using a determiner selection task. In order to test how swapping the systematicity between determiners and suffixes affected individual determiner-picture mapping memory, performance on Experiment 1 was compared with performance on Experiment 3.

As a reminder, there were two determiners in Experiment 1, and four determiners in Experiment 3. To account for this difference, instead of comparing the accuracy proportions statistically, the “difference from chance” was used as the dependent variable. The chance levels differed across the two Experiments. Given that there were two determiners in Experiment 1, the chance level was 0.5. Given that there were four determiners in Experiment 3, the chance level was 0.25.

A two-way between subjects ANOVA was run with the variables of group (sleep/ wake) and language type (Experiment 1/ Experiment 3). There was no significant main effect of language type ($F(1, 100) = 1.65$, $p = .202$), and no significant main effect of group ($F(1, 100) = 1.48$, $p = .226$). There was no significant interaction between language type and group ($F(1, 100) = 1.92$, $p = .169$). As shown in Figure 43, the memory for individual determiner-picture mappings relative to chance was similar in both language types, despite the determiners in Experiment 3 being relatively less systematic than in Experiment 1.

While the performance relative to chance was similar across both Experiments, numerically there was indication that the absolute performance varied. Out of 16 items, in Experiment 3 the average number remembered was 6.46 for the wake group, and 6.23 for the sleep group. Meanwhile, in Experiment 1 the average number remembered out of 16 items was 10.32 for the wake group and 12.25 for the sleep group. Hence, the absolute performance was lower in Experiment 3, than in Experiment 1.

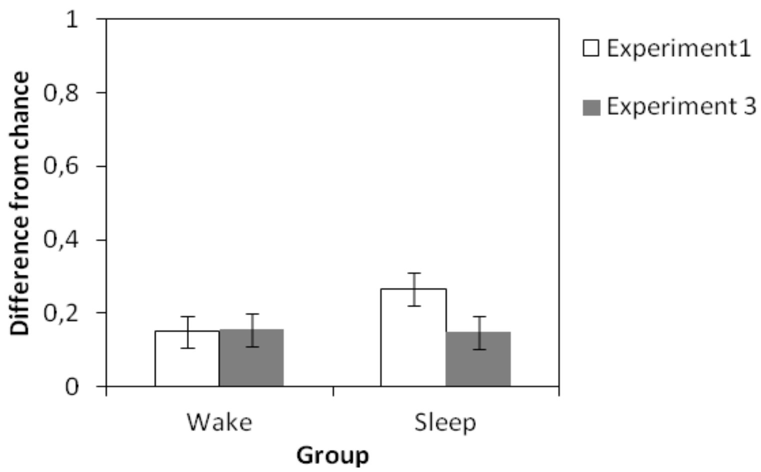


Figure 43. Comparison of performance on the determiner selection task between Experiments 1 and 3.

Determiner and suffix generalisation task

As a reminder, the determiner and suffix generalisation task was used to assess participants’ ability to generalise the trained determiner-gender regularities to previously unseen items, indicative of extraction. To test if any differences between groups in determiner-gender regularity extraction emerged as a result of differences between the two language types, the performance on the determiner and suffix generalisation task was compared between Experiments 1 and 3.

To investigate this, a mixed ANOVA was run. The within subjects variable was consistency (consistent/ inconsistent items). The between subjects variables were group (wake/ sleep) and language type (more systematic determiners/ less systematic determiners). The dependent variable was proportion endorsed. Again, crucially the pattern of endorsement of consistent versus inconsistent items was of interest, as opposed to overall performance.

There was a significant main effect of consistency ($F(1, 100) = 13.60$, $p < .001$). There was no significant main effect of group ($F(1, 100) = .06$, $p = .803$), and no significant interaction between group and consistency ($F(1, 100) = .16$, $p = .691$). There was no significant main effect of language type ($F(1, 100) = .91$, $p = .343$). There was a significant interaction between language type and consistency ($F(1, 100) = 15.32$, $p < .001$). There was no significant interaction between language type and group ($F(1, 100) = 2.06$, $p = .154$). There was no significant three-way interaction between consistency, group and language type ($F(1, 100) = .86$, $p = .357$).

The key findings from this task relate to the pattern of endorsement (i.e. involve consistency), as opposed to overall level of performance, regardless of consistency. As shown in Figure 44, and as reflected by the significant language type x consistency interaction, the pattern of endorsement was in line with determiner-gender regularity extraction in Experiment 1, but not in Experiment 3. There was no evidence that the extraction of determiner-gender regularities differed as a result of group (sleep/wake).

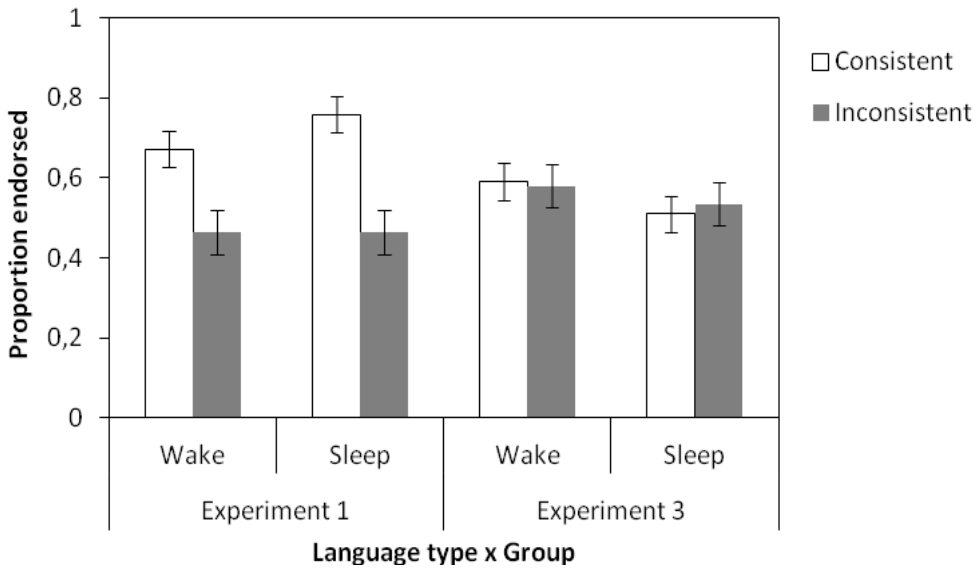


Figure 44. Performance on the determiner and suffix generalisation task between Experiments.

When the determiners were more systematic and the suffixes were less systematic (Experiment 1), participants endorsed consistent items more than inconsistent in this task. When the determiners were made less systematic and the suffixes more systematic (in Experiment 3), endorsement of both consistent and inconsistent items was similar. This suggests that decreasing the systematicity of the determiners and increasing the systematicity of the suffixes may have inhibited determiner-gender regularity extraction.

To summarize the findings from the determiner tests, memory for the individual determiner-picture mappings relative to chance was similar in both Experiments (as measured by the determiner selection task). Numerically, absolute performance was higher in Experiment 1. There was evidence of determiner-picture gender regularity extraction in Experiment 1, but not in Experiment 3. Hence, when determiners were more systematic and suffixes were less systematic, individual determiner memory was higher, and there *was* evidence of determiner-gender regularity extraction. When determiners were less systematic, and suffixes were more systematic, there was *no* evidence of determiner-gender regularity extraction.

Overall summary

Overall, there was no evidence to suggest that decreasing the systematicity of the determiners, and simultaneously increasing the systematicity of the suffixes impacted initial acquisition or recognition memory of the new words. Regardless of the systematicity, there was evidence that sleep benefited the consolidation of the arbitrary mappings (the individual stems and the individual suffixes). There was no evidence to suggest that sleep benefited the extraction of the regularities (the suffixes and the determiners).

For the arbitrary mappings, while there was no evidence that decreasing the systematicity of the determiners and, as importantly, increasing suffix systematicity had an impact on suffix recall, it boosted stem recall. This was potentially because the swapped systematicity made the stems more salient in the input.

For the systematic mappings, there was no *strong* evidence of suffix-gender regularity extraction in either of the Experiments. The patterns of endorsement suggested no evidence for extraction in either case. In addition, when the determiners were more systematic, and the suffixes were less systematic, there was also evidence of determiner-gender regularity extraction. When the systematicity was swapped, there was no evidence of determiner-gender regularity extraction, in the presence of relatively lower individual determiner-picture mapping memory (in terms of absolute performance).

Hence, the two key consequences of decreasing the systematicy of the determiners and simultaneously increasing the systematicity of the suffixes were firstly increased stem recall, and lack of evidence of determiner-gender regularity extraction. Despite suffixes having been made more systematic, there was no evidence that suffix recall was boosted, and no evidence that suffix-gender regularity extraction was boosted. It appears that simplifying suffixes served to boost stem salience (and hence stem recall), as opposed to suffix salience (and hence suffix-gender regularity extraction).

4.3 Discussion

In Experiment 3 (in comparison with Experiment 1 and 2), the systematicity of determiners to suffixes was swapped. Specifically, in comparison with Experiment 1 (and Experiment 2), determiners became less systematic and less frequent in the input (overall 4 determiners, instead of 2). The suffixes became more systematic, and more frequent in the input (overall 2 suffixes, instead of 4). To use the terms of Competition model (e.g. Bates & MacWhinney, 1989), this meant that in Experiment 3, determiners were the less reliable and the less available cue, compared to suffixes.

We found that unlike Experiments 1 and 2, now that the determiners were made less systematic (and hence less frequent in the input), there was no longer any evidence for determiner-gender regularity extraction. Memory for individual determiner-picture mappings, although above chance, was low in terms of absolute performance. However, the extent to which performance deviated from chance was similar between Experiments 1 and 3. Despite suffixes having been made more systematic (and hence more frequent in the input), there was still no clear evidence for the extraction of suffix-gender regularities, consistently with Experiments 1 and 2.

This lack of evidence for *determiner*-gender regularity extraction now that the determiners have been made less systematic is consistent with Kempe and MacWhinney's (1998) findings. Specifically, they found that the less systematic, and less frequently appearing case-markings in German, are more difficult to acquire (in comparison with increased systematicity in Russian). This is also supported by Van den Bos, Christiansen and Misyak (2012), who found that probabilistic mappings are more difficult to acquire than fully systematic mappings. Meanwhile, the lack of extraction of the *suffix*-gender regularity extraction despite the suffixes now being fully systematic requires further exploration (and will be particularly addressed in Chapter 5).

It appears that swapping the systematicity of the regularities had the greatest effect not on the systematic mappings of the suffixes, but on the individual stems at the arbitrary mapping level. While suffix recall was not affected by swapping systematicity, stem recall was boosted by decreasing determiner systematicity and simultaneously increasing suffix systematicity. This might be because the swapping of systematicity made the stems more salient. Consequently, we found some evidence to suggest that this change in systematicity and frequency (i.e. in cue reliability and cue availability to use the Competition model's terms) resulted in greater stem recall (potentially due to increased stem salience).

In terms of sleep effects, in line with findings from Experiments 1 and 2 and predictions of the CLS (e.g. McClelland et al., 1995) we expected to find a benefit of sleep for the consolidation of individual arbitrary mappings, namely on stem recall. Although we did not find a benefit of sleep for stem recall, we did find a benefit of sleep for the consolidation of individual suffixes, as measured by suffix recall. As expected, we found no evidence for a benefit of sleep for the extraction of regularities, consistent with findings from Experiments 1 and 2 and predictions of the CLS (e.g. McClelland et al., 1995).

Experiment 4 was an exploratory investigation which addressed two separate questions. The first question asked whether additional training would boost determiner-gender regularity and suffix-gender regularity extraction. It was also of interest whether differences in wake-dependent consolidation versus sleep-dependent consolidation on performance on language tasks measuring arbitrary and systematic mappings would emerge as a result of additional training. The second question asked whether determiner-gender regularities and suffix-gender regularities are available immediately. An immediate group was also included in Experiment 4 to address this question.

Initial Level of training

The findings from Experiment 3 did not provide any evidence of suffix-gender regularity or determiner-gender regularity extraction, when determiners were less systematic than suffixes in the input. In order to understand why there was no extraction of the more systematic suffixes, Experiment 4 seeks to provide additional training to strengthen initial encoding and initial memory traces. The memory trace needs to be available initially, in order for sleep to be able to act on it to extract regularities. This can be illustrated in the "information overlap to abstract" ("iOtA") model (Lewis & Durrant, 2011), as discussed in Chapter 1. This model suggests that when information is learnt which contains within it overlapping elements, during sleep memories are reorganised in such a way that the overlapping elements are strengthened while the specific individual items are eroded. The strengthened memory trace for the overlapping elements benefits their abstraction following sleep. Hence, for sleep to be able to re-organise the information to boost extraction, the overlapping elements need to be sufficiently encoded in the first place. Indeed, individual suffix recall was found to be low in Experiments 1, 2 and 3, when suffix-gender regularities were not extracted.

Providing extra training would strengthen the initially encoded traces. This is important, as given that we are predicting (based on the CLS, McClelland et al., 1995), that sleep does not benefit extraction of regularities, in order to be more confident in our results, we need to investigate any other factors which might be contributing to the role of sleep in extraction. The strength of the initial trace is one such factor.

However, currently there is conflicting evidence within existing literature, whether stronger or weaker initial memory traces may result in a greater benefit of sleep.

Specifically, for example, Talamini, Nieuwenhuis, Takashima, and Jensen (2008) found that initially stronger memory traces, compared to initially weaker memory traces, have significantly stronger resistance to forgetting following sleep. This was found using a face location task, whereby participants learned 20 different face-location pairs. Participants saw 8 different grey circles on screen. On each trial, a photograph of a different face appeared in the centre. One of the grey circles turned green, signalling the target location. Participants moved their cursor to the target location, and consequently the face re-appeared at the target location. Participants' recall of the face-location pairs was tested both immediately and also after either 12 hours (containing sleep or wakefulness), or after 24 hours. It was found that the benefit of sleep for this episodic type of memory depended on initial performance. Sleep benefited the initially stronger memories, but not the initially weaker memories.

In contrast, are the findings of Drosopoulos, Schulze, Fischer and Born (2007). They taught participants one set of word-pair associates, later followed by a second set of word-pair associates used to create retroactive interference. Initial learning performance was compared with performance either after a night of sleep in the lab, or after the equivalent time of daytime wakefulness. Participants from both groups had similar performance after the consolidation delay, on the pairs which were initially encoded well. Participants who slept overnight had better performance (compared to the participants from the daytime wakefulness group) on the pairs which were initially encoded less well. The findings suggested that sleep benefits the memory for items which are encoded weakly.

To summarize, in order for mappings to be extracted, the memory traces for mappings should be available initially. Additional training should strengthen the existing memory traces. Therefore, if suffix-gender regularities were not extracted due to not having sufficiently strong mappings in Experiment 3, additional training should boost suffix-gender regularity extraction in Experiment 4 (i.e. in this Experiment). The same may also apply to the determiners, as there was no evidence of determiner-gender regularity extraction in Experiment 3. The absolute performance on the determiner selection task indicated low individual determiner-picture mapping memory in Experiment 3. Absolute performance in Experiment 3 was also numerically lower than in Experiment 1, when determiner-gender regularities were extracted.

In order to investigate this idea that increased training might boost regularity extraction, the word-picture matching task which was used to train participants on the new form-meaning mappings was increased in length. In Experiment 4, this task provided participants with twice as many repetitions of each mapping as in Experiments 1-3. The key question was whether additional training will influence performance on the task measuring arbitrary and systematic mappings, and whether this performance will differ as a result of wake-dependent consolidation compared to sleep-dependent consolidation. It was predicted that providing additional training should provide more opportunity for regularities to be extracted, given stronger memory traces.

Immediate memory

Experiments 1 -3 have not shown any evidence for suffix-gender regularity extraction after a consolidation delay. Additionally, in Experiment 3 there was no evidence of the extraction of the determiner-gender regularity. As with Experiments 1-3, the main focus in Experiment 4 is still on the distinction of the role of sleep and wake for the consolidation of arbitrary mappings, and the extraction of systematic mappings. However, it would also be useful to have information about the immediate memory of both arbitrary and systematic mappings. It is of interest whether the determiner-gender regularities and suffix-gender regularities are available immediately.

Hence, an additional set of participants was tested immediately to provide this information. Half of the participants in this immediate group were trained and tested in the morning. The other half were trained and tested in the evening. This was done in order to account for circadian effects.

This is an exploratory investigation, which will investigate whether regularities are extracted immediately or not. There are three possible outcomes for both the determiners (given that there was no extraction of the determiner-gender regularity in Experiment 3, unlike Experiments 1 and 2) and the suffixes. In one possible outcome, there may be no evidence of extraction (this applies to both determiners and suffixes) after a consolidation delay, but evidence of extraction immediately. This would suggest that the regularity is extracted initially, but that the knowledge of the regularity is lost (forgotten) after a consolidation delay. In a second possible outcome, there may be evidence of extraction after a delay, but not immediately. This would suggest that a period of consolidation is required for extraction of the regularity to emerge. In a final possible outcome, there may be no evidence of extraction – neither after a consolidation delay, nor immediately. This might suggest that the lack of regularity extraction cannot be explained by factors relating to consolidation processes. Instead, it might be the result of initial acquisition problems, due to factors within the language input. Such factors include stem variability, and will be considered in more detail in the discussion.

4.4.1 Method

It is important to note that fatigue is a potential confound in this study. To elaborate, to test whether extraction was available immediately, the immediate group was tested immediately after training: there was only a short 15 minute break between the training session (session 1) and the test session (session 2). As with Experiments 1 and 3, in Experiment 4 (i.e. in this Experiment), the two delayed groups had 12 hours' break between sessions 1 and 2. Hence, the immediate group may have had more fatigue during the session 2 language tests, compared to the two delayed groups. We cannot directly address circadian factors, as they were confounded with fatigue at session two. In order to account for circadian factors, half of the immediate group was trained and tested in the morning; the remaining half was trained and tested in the evening.

A key question in Experiment 4 is whether there was emergence of regularity extraction (of determiner-gender and suffix-gender mappings) initially, that is forgotten following a delay. Hence, if there is evidence of extraction immediately (of either determiner-gender regularities or suffix-gender regularities), it will not be as a result of circadian factors, given that these have been accounted for. It is reminded that the primary question of this investigation is the importance of the initial level of acquisition for extraction, as opposed to testing circadian factors.

The *training* performance of the two delayed groups (i.e. the wake and sleep groups) will be compared statistically against the immediate group. As the AM-immediate group was trained at the same time as the wake group, and the PM-immediate group was trained at the same time as the sleep group, and no other differences had yet transpired between the groups, session one performance was comparable. Session two analyses will focus on qualitative comparisons, i.e. testing to see if there is extraction of regularities via the main effect of consistency. This is because session two performance is likely to be more influenced by fatigue in the immediate groups (as shown in the SSS results in the appendices).

Participants

The participant criteria were the same as in Experiments 1, 2 and 3. As with Experiment 1 and 3, there were 2 delayed groups in Experiment 4: the sleep group and the wake group. There was an additional immediate group. Half of the immediate group was trained and tested in the morning; the other half was trained and tested in the evening.

There were a total of 106 participants in Experiment 4. Two participants did not complete the second session of the study and therefore were excluded from the analysis. Out of the remaining 104 participants, twenty-six participants with a mean age of 19.58 (± 1.92 [SD]) were in the sleep group (5 male, 3 left handed). Twenty-seven participants with a mean age of 23.59 (± 7.64 [SD]) were in the wake group (8 male, 1 left handed). Twenty-six participants with a mean age of 19.31 (±1.32 [SD] were in the immediate AM-tested group (8 male, 2 left handed). Twenty-five participants with a mean age of 18.84 (± 0.90 [SD]) were in the immediate PM-tested group (4 male, 1 left handed).

Out of those participants, there were 8 participants in the sleep group, 15 participants in the wake group, 10 participants in the immediate AM-tested group, and 12 participants in the immediate PM-tested group, who played either a string instrument or the piano (according to answers on session 1 questionnaire). In return for participation, £18 or 3 hours undergraduate psychology course credit was given, due to the extended training time.

Design

The design was the same as in Experiment 3 with one key difference. An additional group of participants was tested immediately, with a 15 minute break between session 1 and session 2 (as opposed to a 12 hour consolidation delay). Half of the participants were trained and tested in the morning, the remaining half were trained and tested in the evening, in order to account for circadian effects.

During the 15 minute break, the participants were required to leave the experimental room. All of the participants then returned and completed all of the session 2 tests in the same order as the delayed groups.

Materials

The materials were the same as in Experiment 3, with the exception of an additional period of training.

An additional period of training was included for all participants, increasing the training session from 60 to 100 minutes. There were now 384 trials in total, as opposed to 192 as in Experiment 3. Each 1 of the 16 target word-picture combinations was seen a total of 16 times (instead of 8). Each word was seen with 8 different (instead of 4 different) noise picture combinations, out of which half were masculine and half were feminine. This means that the ratio of target to noise trials was kept constant between Experiment 3 and 4. The 384 trials were separated into 8 blocks, with 2 target trials and 1 noise trial for each item within the block.

Procedure

The procedure and order of tasks was the same as in Experiment 3. The key difference was that participants from the two immediate groups had a 15 minute break between sessions 1 and 2.

4.4.2 Results

Language tasks

Training task (session one)

As with Experiments 1-3, participants were trained on the artificial language using a word-picture matching task. Also in line with Experiments 1- 3, there were two language versions, and both an accuracy analysis and an RT analysis were included to provide an indication of the learning of the language.

Accuracy analysis

Participants' accuracy at the end of training was measured to assess their knowledge of the mappings and compared between the experimental groups (sleep and wake as with Experiment 3, and additionally the immediate group) and the two language versions. As this was the participants' first exposure to the language, it was expected that there would be no significant differences between groups.

A two way between subjects ANOVA was run with the independent variables of group (sleep, wake or immediate) and language (version one or two). The dependent variable was accuracy (proportion of correct responses for the "target" trials only) at the final block of the training task. Given the unequal sample sizes, Hochberg's GT2 was used for post hoc multiple comparisons. There was a significant main effect of group ($F(2, 98) = 8.45$, $p < .001$). There was no significant main effect of language version ($F(2, 98) =

.41, p = .525) and no significant interaction between group and language ($F(2, 98) = 1.03$, p = .362).

According to the Hochberg's GT2, there was a significant difference in performance between the wake and the sleep group ($p < .001$), a significant difference in performance between the immediate and the sleep group (p = .009), and no significant difference in performance between the immediate and the wake group (p = .321).

As shown in Figure 45, by the end of training the wake group reached a higher level of accuracy than the sleep group. Similarly, the immediate group also reached a higher level of accuracy than the sleep group. There was no evidence to suggest that language version affected performance.

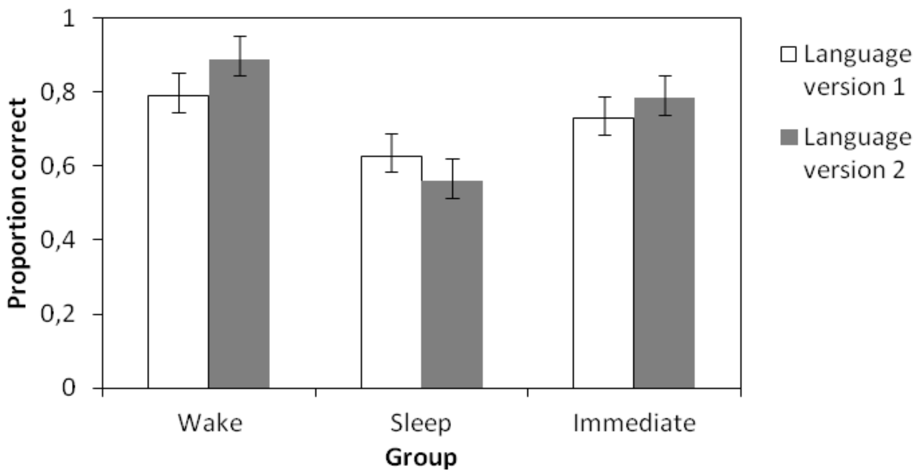


Figure 45. Accuracy at the end of the training task for the immediate and delayed groups.

Correct RTs analysis on the training task

As with Experiments 1-3, an analysis of the RTs was run. It was expected that participants' RTs should speed up with learning. Instead of taking raw RTs, the RTs from the first block of training were subtracted from each of blocks 2 to 8. This was done for correct responses to target RTs, and separated for correct responses to noise RTs, as with previous Experiments. Because of the difference between groups, the Hochberg's GT2 was applied.

A mixed ANOVA was run, with the within subjects variable of block difference (difference between blocks 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, 8 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake, sleep and immediate). The dependent variable was the average correct target RTs.

Three participants (2 from the wake group, 1 from the sleep group) were excluded from the analysis, as they did not have any correct responses for target items in some of the blocks. There was a significant main effect of block ($F(2, 95) = 42.32$, $p < .001$). There was no significant main effect of group ($F(2, 95) = 1.91$, $p = .154$). There was a significant interaction between block and group ($F(2, 95) = 2.80$, $p < .001$). The Hochberg's GT2 posthoc confirmed that there were no significant differences between any of the groups. More specifically, there were no significant differences between the wake and the sleep group ($p = .237$), no significant differences between the wake and the immediate group ($p = .218$), and no significant differences between the sleep and the immediate group ($p = .994$), in overall performance.

There was no significant main effect of language version ($F(2, 95) = .13$, $p = .719$), and no significant interaction between group and language version ($F(2, 95) = .62$, $p = .543$). There were no significant interactions between block and language version ($F(2, 95) = .82$, $p = .552$), or block, group and language ($F(2, 95) = .98$, $p = .464$). As shown in Figure 46, while all three groups got faster with training, the difference in RTs compared to block 1 was greatest for the wake group. As suggested by the significant group x block interaction, the wake group appears to have improved more than the sleep group and the immediate group. To account for this difference, the target RT difference between the final block of training (block 8) and block 1 will be included as a covariate in the session two language task analyses.

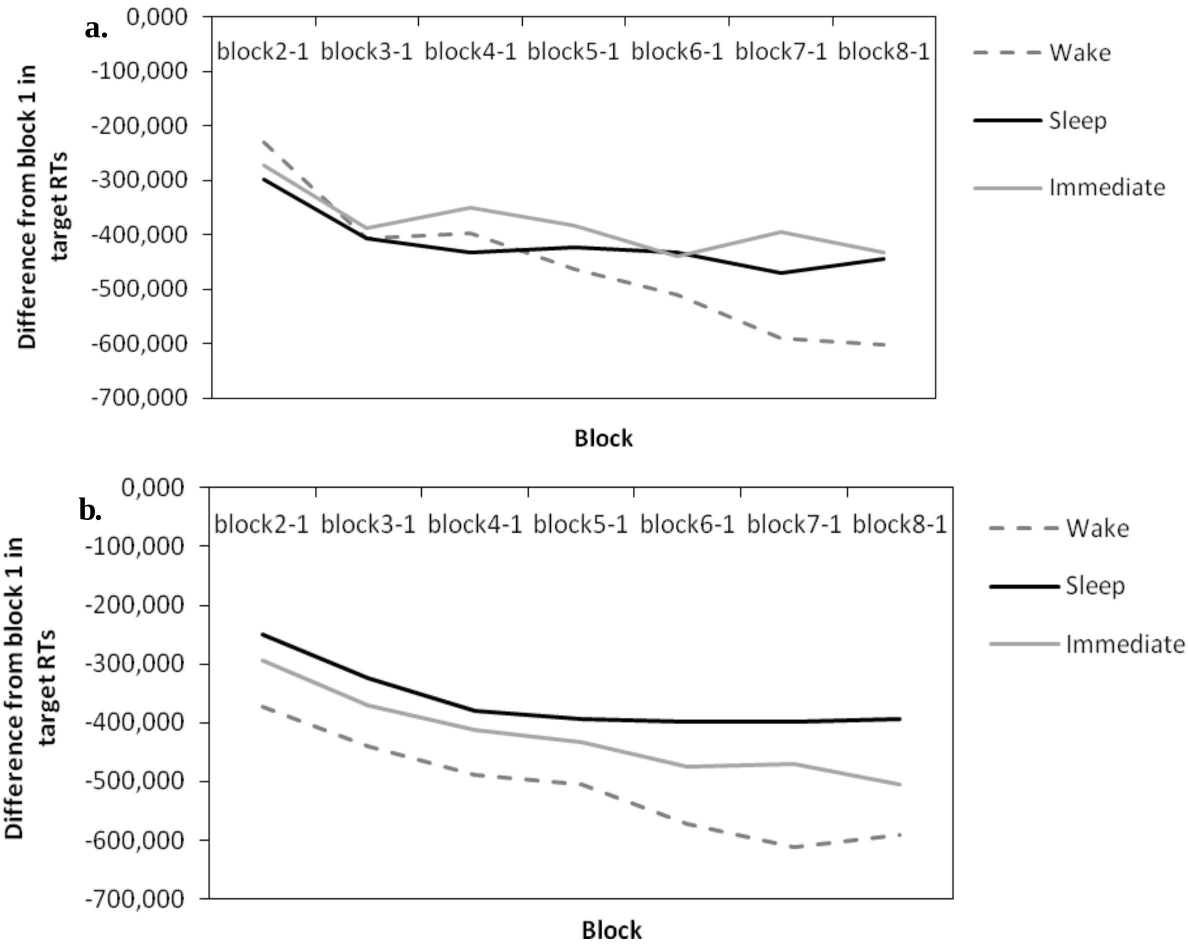


Figure 46. Difference in *target* RTs from block 1 for the immediate group and the delayed groups (Figure 46a. shows language version 1 RTs, Figure 46b. shows language version 2 RTs).

Secondly, the dependent variable was the change in noise RTs compared to block 1. Two participants (1 participant from the wake group and 1 participant from the immediate group) could not be included in the analysis, as they did not have any correct responses to noise trials in all of the blocks. Because of the difference between groups, the Hochberg's GT2 was applied.

There was a significant main effect of block ($F(2, 96) = 35.58$, $p < .001$). There was no significant main effect of group ($F(2, 96) = .80$, $p = .452$), and no significant interaction between block and group ($F(2, 96) = 1.55$, $p = .104$). The Hochberg's GT2 posthoc confirmed that there were no significant differences between any of the groups. More specifically, there were no significant differences between the wake and the sleep group ($p = .933$), no significant differences between the wake and the immediate group (p

= .507), and no significant differences between the sleep and the immediate group (p = .887).

There was no significant main effect of language version ($F(2, 96) = .11$, $p = .739$) and no significant interaction between group and language ($F(2, 96) = .38$, $p = .683$). There were also no significant interactions block and language ($F(2, 96) = 1.19$, $p = .312$), or block, group and language ($F(2, 96) = 1.03$, $p = .420$). As shown in Figure 47, the difference in RTs compared to block 1 increased with training, regardless of group.

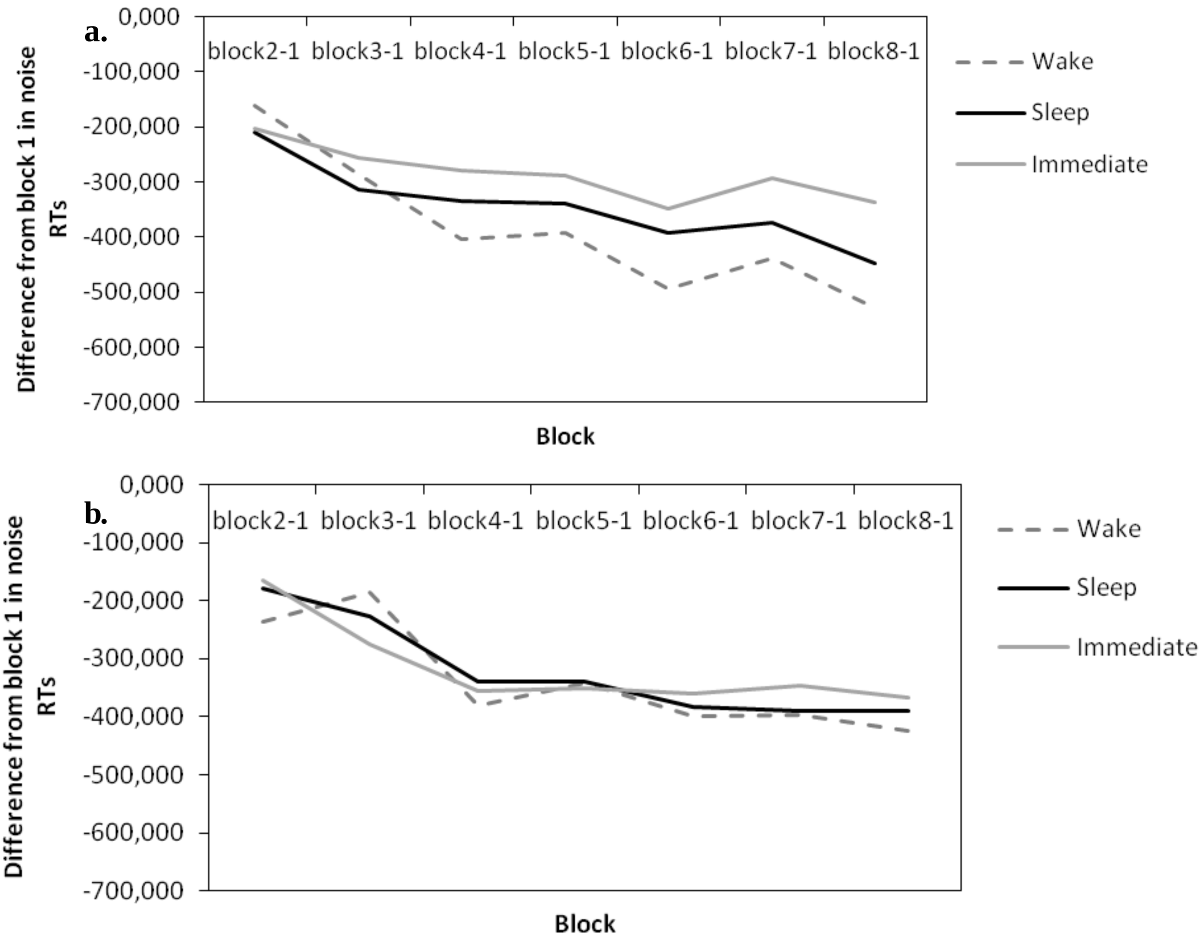


Figure 47. Difference in *noise* RTs from block 1 for the immediate group and the two delayed groups (Figure 47a. shows language version 1 RTs, Figure 47b. shows language version 2 RTs).

To summarize, the results from the training task accuracy analysis have suggested that the wake and the immediate groups may have improved more than the sleep group.

Consequently, session 1 training performance (more specifically, the difference between block 8 and block 1 for correct target RTs) will be included as a covariate in session two tests. As both language versions had a similar level of performance, all other tasks will be analysed collapsing across language version.

Language tests (session two)

As outlined in the introduction to Experiment 4, there were two research questions that were investigated in Experiment 4. Firstly, it was of interest whether additional training influences performance across the tasks (i.e. for the arbitrary and systematic mappings). Secondly, it was of interest whether the performance across tasks changes when participants are tested immediately. To address these two questions, the results will be presented in two separate subsections. The first subsection will address whether additional training made a difference for delayed consolidation (i.e. sleep in comparison to wake-dependent consolidation). The delayed data was analysed with 25 participants in the sleep group and 25 participants in the wake group. As the training performance was included as a covariate, all figures are presented on the basis of the adjusted means.

The second subsection will address whether the pattern of performance across tasks is different when participants are tested immediately. As already explained in the introduction, as a consequence of the fatigue confound at session two (see Appendix), the analyses for the immediate group (i.e. in the second section) will focus on qualitative comparisons of the immediate group.

Given that the crucial question is whether extraction of suffix-gender regularity and determiner-gender regularity is available immediately, but forgotten with a delay, the key analyses are concerned with the main effects of consistency in the two generalisation tasks.

4.4.2.1 Results of the delayed groups

Most arbitrary mapping: vocabulary

Translation recognition task

As in Experiments 1-3, participants' memory for the trained words and their meanings was tested using a translation recognition task. Half of the trials contained the match (correct) English word-novel word pairing translations, half contained mismatch (incorrect) pairings. As with Experiments 1-3, to account for the potential differences in participants' ability to discriminate between correct and incorrect pairings, the non-parametric (A') discrimination formula (Donaldson, 1992) was used to analyse the data from the delayed groups.

A one-way ANCOVA was run with the independent variable of group (sleep or wake), training performance as the covariate, and A' discrimination as the dependent variable. There was no significant main effect of group ($F(1, 46) = 1.65$, $p = .206$). There was no significant main effect of training as the covariate ($F(1, 46) = .47$, $p = .495$), and no interaction between group and training as the covariate ($F(1, 46) = .87$, $p = .354$). Participants from both of the delayed groups had similarly good performance on the translation recognition task, as shown in Figure 48. For A', 0.5 indicates that match trials cannot be distinguished from mismatch trials, and 1.0 indicates perfect discrimination. Accordingly, all participants across the two groups had similar discrimination between correct and incorrect trials.

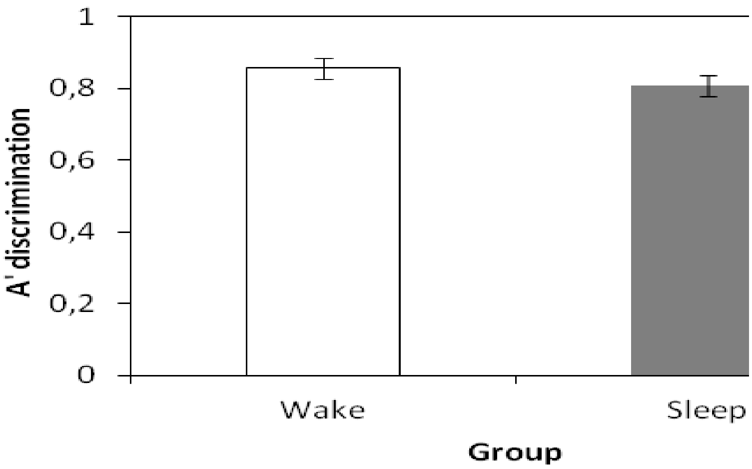


Figure 48. Performance of the delayed groups on the translation recognition task.

Recall (stems)

As with Experiments 1-3, participants' memory for the trained word-picture mappings was tested using a picture naming recall task. In line with Experiments 1-3, data from this task was analysed for stems and suffixes separately. To analyse stem recall for the delayed groups, a two-way mixed ANCOVA was run. The between subjects variable was group (wake/sleep), the within-subjects variable was cue type. Training performance was included as the covariate and stem recall accuracy (proportion correct out of total productions) was the dependent variable.

There was no significant main effect of group ($F(1, 46) = .69$, $p = .411$) and no significant main effect of cue type ($F(1, 46) = .84$, $p = .364$). There was a significant interaction between group and cue type ($F(1, 46) = 4.98$, $p = .031$). As shown in Figure 49, participants from the wake group benefited from the grapheme cue more than participants from the sleep group. Although numerically the performance of participants in the sleep group was overall higher than the performance of participants in the wake group, this was not significant.

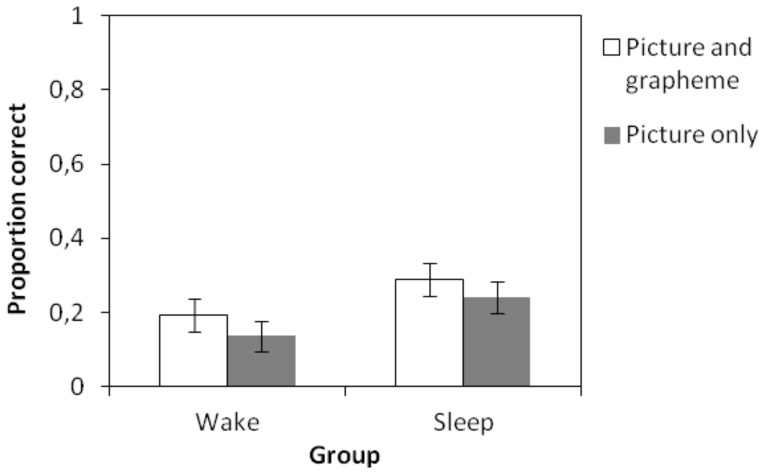


Figure 49. Stem recall accuracy for the delayed groups.

Additionally, there was no significant effect of training as the covariate ($F(1, 46) = .36$, $p = .554$), no significant training x group interaction ($F(1, 46) = .03$, $p = .865$), no significant training x cue type interaction ($F(1, 46) = .03$, $p = .869$). There was a significant three way interaction between group x training x cue type ($F(1, 46) = 5.41$, $p = .024$).

This three-way interaction can be unpacked by looking at the means. The means for the wake group *without* taking training into account were .176 for the picture and grapheme, .141 for the picture only. The means for the wake group taking training into account were .193 for the picture and grapheme, .136 for the picture only. The means for the sleep group *without* taking training into account were .285 for the picture and grapheme, .250 for the picture only. The means for the sleep group taking training into account were .289 for the picture and grapheme, .241 for the picture only. This suggests that training affected the wake group's performance when the grapheme was presented more than when the picture only was present. It was vice versa for the sleep group, performance was affected more when the picture only was present, than when the grapheme was present as well.

Overall, the results from the vocabulary tests suggest that all participants had a similarly good level of recognition memory as evidenced by the translation recognition task. Unlike Experiments 1 and 2, there was no evidence for a benefit of sleep over wake on individual stem recall.

Most systematic mapping: suffixes (grammar)

For the suffix tests, we were interested in whether any differences as a result of group (sleep/wake) emerged in the consolidation of the suffix-picture mappings and in the extraction of suffix-gender regularities. This was measured by the suffix recall task, and by the suffix generalisation task, and more specifically by looking at the main effect of consistency on the suffix generalisation task as indication of suffix-gender regularity extraction.

Recall (suffixes)

As with Experiments 1-3, to assess participants' memory of the systematic element of the word, the recall accuracy of the suffixes (which were separated from stems in the transcriptions) was analysed.

For the delayed groups, a two-way mixed ANCOVA was run. The between subjects variable was group (wake/sleep), the within-subjects variable was cue type (picture and grapheme, or picture only). Training performance was included as the covariate. Suffix recall accuracy (proportion correct out of total productions) was the dependent variable. There was no significant main effect of group ($F(1, 46) = .48$, $p = .494$) and no significant main effect of cue type ($F(1, 46) = .01$, $p = .935$). There was no significant interaction between group and cue type ($F(1, 46) = .07$, $p = .801$).

Additionally, there was no significant effect of training as the covariate ($F(1, 46) = .06$, $p = .814$), no significant training x group interaction ($F(1, 46) = .02$, $p = .886$), no significant training x cue type interaction ($F(1, 46) = 1.13$, $p = .293$). There was no significant three way interaction between group x training x cue type ($F(1, 46) = .26$, $p = .610$). Consequently, there was no evidence that training performance affected suffix-picture mapping consolidation.

As shown in Figure 50, although numerically participants in the sleep group had overall higher suffix recall accuracy than participants in the wake group, there was no statistical evidence that sleep benefited the consolidation of the trained suffix-picture mappings.

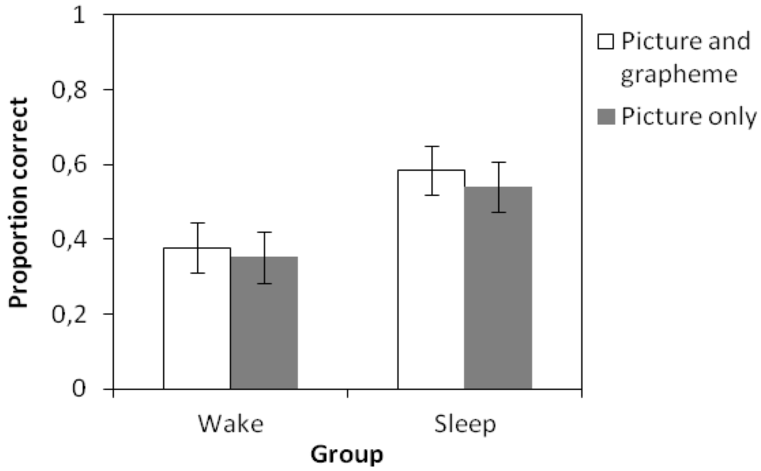


Figure 50. Suffix accuracy on the recall task for the delayed groups.

Suffix generalisation task

As with Experiments 1-3, to test if participants extracted the suffix-gender regularities which they were exposed to during training, a suffix generalisation task was used. Crucially, the pattern of endorsement of consistent versus inconsistent trials was of interest here, as opposed to the overall level of performance.

For the delayed groups, a two-way mixed ANCOVA was run. The between subjects variable was group (wake/sleep), the within-subjects variable was consistency (items consistent versus inconsistent with the trained regularities). Training performance was included as the covariate. Proportion endorsed was the dependent variable. There was no significant main effect of group ($F(1, 46) = .87$, $p = .356$) and no significant main effect of consistency ($F(1, 46) = .00$, $p = .963$). There was no significant interaction between group and consistency ($F(1, 46) = .37$, $p = .545$). As shown in Figure 51, participants had a similar level of endorsement regardless of consistency; hence there was no evidence of suffix-gender regularity extraction, regardless of group.

Additionally, there was no significant effect of training as the covariate ($F(1, 46) = .12$, $p = .727$), no significant training x group interaction ($F(1, 46) = 2.05$, $p = .159$), and no significant training x consistency interaction ($F(1, 46) = .02$, $p = .885$). There was also no significant three way interaction between group x training x consistency ($F(1, 46) =$

.09, $p = .761$). Consequently, there was no evidence that training performance affected suffix-gender regularity extraction.

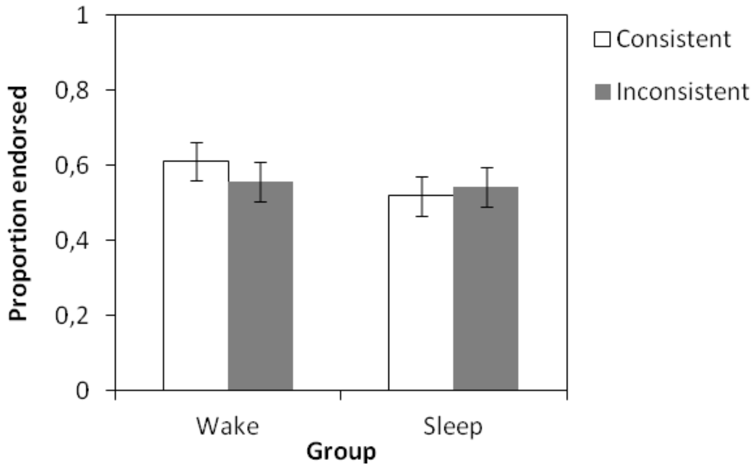


Figure 51. Suffix generalisation performance for the delayed groups.

As there was no main effect of consistency, and no interaction between group and consistency, there was no evidence of suffix-gender regularity extraction regardless of group, and no evidence of group differences in suffix-gender regularity extraction. To investigate this further (as with Experiments 1-3), one sample t-tests against chance were run, separately for consistent and inconsistent items.

The results of one-sample t-tests against chance (0.5) are shown in Table 10. Participants from the sleep group endorsed both consistent *and* inconsistent items at chance. Participants from the wake group endorsed consistent items significantly above chance, and were at chance at endorsing inconsistent items. This pattern of endorsement in the wake group *is* in line with the pattern expected for suffix-gender regularity extraction. Hence, there was some evidence from the one sample t-tests against chance to suggest that the wake group only, and *not* the sleep group was able to extract suffix-gender regularities.

Table 10. Results of one sample t-tests against chance on the suffix generalisation task.

	Word-picture gender pairing	t	df value	*p* - value
sleep	Consistent	.88	24	.388
	Inconsistent	1.54	24	.136
wake	Consistent	2.70	24	.012
	Inconsistent	1.41	24	.170

To summarize, the findings from the suffix tests suggest that participants from the delayed groups reached a similar level of suffix recall accuracy, regardless of sleep. Consequently, unlike Experiments 1-3, there was no evidence for a benefit of sleep for individual suffix recall. Although there was no *strong* evidence of suffix-gender regularity extraction after a consolidation delay (due to the lack of a main effect of consistency), based on the one sample t-tests against chance, there was *some* evidence to indicate extraction of suffix-gender regularities in the wake group only.

It might be *tentativel*y suggested that suffix-gender regularity extraction may start to emerge with wake-dependent consolidation, as opposed to sleep-dependent consolidation, if suffix-gender regularity is not available immediately. Alternatively, if suffix-gender regularity extraction *is* available immediately, wake-dependent consolidation may protect this extraction from being forgotten. The analysis of performance of the immediate group (presented later, in a separate subsection as already mentioned) will further confirm which of these two tentative suggestions may be correct.

Intermediate systematicity mapping: determiners (grammar)

Determiner selection task

As with Experiments 1-3, participants' memory for the trained determiner-picture mappings was assessed using a determiner selection task. For the delayed groups, a one-way ANCOVA was run with the independent variable of group (sleep or wake), training performance as the covariate, and proportion correct as the dependent variable. There was no significant main effect of group ($F(1, 46) = 1.26$, $p = .267$). There was no significant main effect of training as the covariate ($F(1, 46) = .01$, $p = .946$), and no interaction between group and training as the covariate ($F(1, 46) = .58$, $p = .449$).

As shown in Figure 52, both of the delayed groups reached a similarly low level of accuracy. The suggestion that memory for individual determiner-picture mappings was low is further supported by looking at absolute performance. The average accuracy count out of 16 items was 6.5 for the sleep, and 5.81 for the wake group.

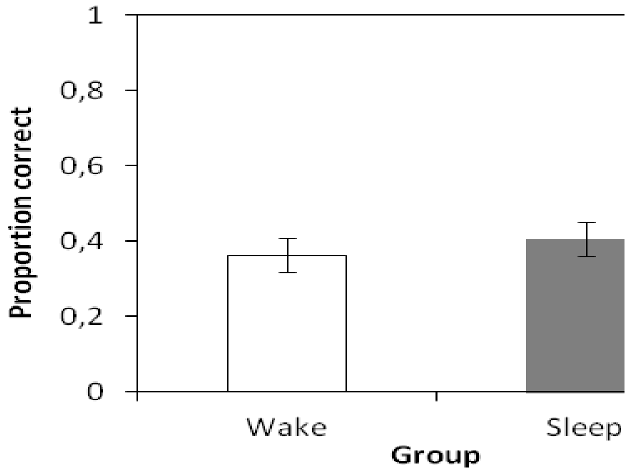


Figure 52. Performance on the determiner selection task.

The results of one-sample t-tests against chance (set at 0.25, given that there were 4 determiners), are shown in Table 11. The results indicated that while participants had a low level of accuracy, all participants were above chance at correctly selecting determiner-picture mappings. This suggests that all participants had memory traces available for the individual determiner-picture mappings, from which the determiner-gender regularities could potentially be extracted.

Table 11. Results of Experiment 4 one sample t-tests against chance on the determiner selection task.

	t	df value	*p* - value
Sleep	4.09	24	< .001
Wake	2.57	24	.017

Analysis of incorrect responses on the determiner selection task

As with Experiment 3, to investigate whether participants may have relied on the extracted determiner-gender regularity to complete the determiner selection task, participants' errors were analysed. More specifically, errors reflecting correct determiner-gender mappings (proportion of correct gender determiners out of the incorrect but trained determiners) were compared against the chance level of 0.33. This is because given that there were four determiners in the input, each response could have three possible incorrect trained determiners. Given that there were two determiners per gender, only one of the three could be of the correct gender (as with Experiment 3). Relying on the extraction of the determiner to picture gender mapping to complete the task would be reflected by correct determiner-gender mappings being produced above chance.

As shown in Table 12, all of the participants were at chance at selecting the correct determiner-gender mapping when making an error. This suggests that not only participants' individual item memory for determiner-picture mappings was low, but participants also did not show any evidence of relying on the determiner-gender *regularity* to complete this task. This is in line with the analysis from Experiment 3 (also using 4 determiners).

Table 12. Results of correct determiner-gender errors analysis in the determiner selection task using one sample t-tests.

Group	t	df value	*p-value*	Mean proportion
Sleep	-1.0	24	.704	.32
Wake	-.38	24	.329	.31

Determiner and suffix generalisation task

As with Experiments 1-3, as an additional test of the extraction of the trained determiner-gender regularities, the determiner and suffix generalisation task was used. As a reminder, in this task we were interested in the pattern of endorsement, as opposed to overall performance.

For the delayed groups, a two-way mixed ANCOVA was run. The between subjects variable was group (wake/sleep), the within-subjects variable was consistency (items consistent versus inconsistent with the trained regularities). Training performance was included as the covariate. Proportion endorsed was the dependent variable. There was no significant main effect of group ($F(1, 46) = .445$, $p = .508$) and no significant main effect of consistency ($F(1, 46) = .06$, $p = .811$). There was no significant interaction between group and consistency ($F(1, 46) = .33$, $p = .566$). As shown in Figure 53, both of the delayed groups had a similar level of endorsement regardless of the consistency of the items. Hence, there was no evidence of determiner-gender regularity extraction in the two delayed groups, and no additional benefit of sleep.

Additionally, there was a marginally significant effect of training as the covariate ($F(1, 46) = 3.56$, $p = .065$), no significant training x group interaction ($F(1, 46) = .76$, $p = .388$), no significant training x consistency interaction ($F(1, 46) = .07$, $p = .782$). There was no significant three way interaction between group x training x consistency ($F(1, 46) = .03$, $p = .868$).

The key finding in terms of the training covariate would be the training x consistency interaction (without or without group). This is because the overall level of performance was not informative about extraction. Rather, the pattern of endorsement according to consistency was informative. Crucially, there was no training x consistency interaction, and no three-way training x consistency x group interaction. Hence, there was no evidence to suggest that training affected extraction of determiner-gender regularities. Less importantly, the overall endorsement mean, regardless of consistency, without taking training into account was .574 and taking training into account was .564. Hence, overall rate of endorsement (regardless of consistency) was lower when training was taken into account.

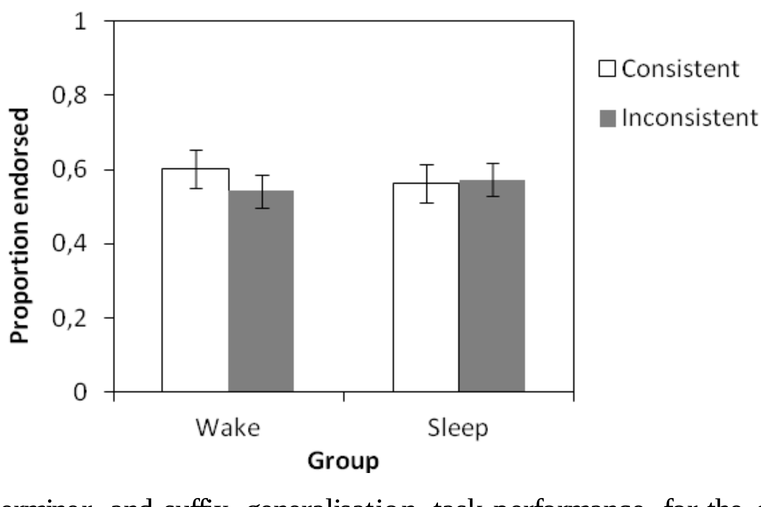


Figure 53. Determiner and suffix generalisation task performance for the delayed groups.

As with Experiments 1-3, performance was additionally compared against the chance level of 0.5. As a reminder, if determiner-gender regularities are extracted, the expected pattern is that participants would be significantly above chance on the consistent items, and either at chance or significantly below chance on the inconsistent items. As shown in Table 13, participants in the sleep group were marginally above chance on the consistent items, and significantly above chance on the inconsistent items. Participants in the wake group were significantly above chance on the consistent items and at chance on the inconsistent items. Hence, there was some indication that the wake group but not the sleep group displayed the pattern that would be expected if the determiner-gender regularities were extracted, although there was no group x consistency interaction to support this.

Table 13. Results of one-sample t-tests against chance on the determiner and suffix generalisation task.

	Word-picture gender pairing	t	df value	*p* - value
sleep	Consistent	1.89	24	.071
	Inconsistent	2.20	24	.038
wake	Consistent	2.43	24	.023
	Inconsistent	.231	24	.819

To summarise the findings from the determiner tests, the determiner selection task analysis provided evidence that memory for individual determiner-picture mappings while above chance, was low. The analysis of the errors on the determiner selection task did not provide any evidence that participants relied on determiner-gender regularity extraction to complete the task, in either of the two delayed groups.

In the determiner and suffix generalisation task, while there was no *strong* evidence of determiner-gender regularity extraction in the delayed groups, the tests of performance against chance provided *some* indication that the wake group was showing some evidence of extraction of the determiner-gender regularities. There was no evidence of determiner-gender regularity extraction in the sleep group. However, although the one sample t-tests against chance indicated that the wake group was showing evidence of determiner-gender regularity extraction, this was not supported by the findings from the analysis of errors on the determiner selection task. In addition, there was *no* interaction between consistency and group on the determiner and suffix generalisation task, that would have indicated that the wake group was better than the sleep group at extracting the determiner-gender regularity.

Overall summary

In Experiment 4, we increased the level of training. It was of interest whether as a result of this increased training, extraction of determiner-gender regularities and suffix-gender regularities would emerge.

For suffixes, there was no *strong* evidence of suffix-gender regularity extraction, in either of the two delayed groups. There was however *some* evidence from the one sample t-tests against chance to indicate the extraction of the suffix-gender regularities in the wake group only, but not in the sleep group.

For determiners, there was also no *strong* evidence of determiner-gender regularity extraction. There was *some* indication that the wake group was showing extraction of the determiner-gender regularity, based on the results of the one-sample t-tests against chance on the determiner and suffix generalisation task. However, this was *not* supported by findings from the analysis of errors on the determiner selection task. There was also *no* group x consistency interaction on the generalisation task that would have indicated that the performance of the wake group was more in line with determiner-gender regularity than the performance of the sleep group.

Consequently, there was *some* evidence that the wake group was showing extraction of determiner-gender regularities and suffix-gender regularities. It is of interest whether this extraction (of both determiner-gender regularities and suffix-gender regularities) emerged as a result of wake-dependent consolidation (if not available immediately). Alternatively, it could be that this extraction *was* available immediately, but consequently forgotten over the consolidation delay, with slightly less forgetting following wake-dependent consolidation, than following sleep-dependent consolidation. The analysis of the immediate group's performance (presented in the next section) will provide evidence as to which of the two suggestions may be correct.

For the arbitrary measures, there was a similarly good level of recognition memory for all participants, with no additional benefit of sleep. There was no evidence of a benefit of sleep in neither stem recall nor suffix recall, unlike Experiments 1and 2.

4.4.2.2 Results of the immediate group

Language tests (Session 2)

It was of interest whether determiner-gender regularities and suffix-gender regularities were extracted immediately. Consequently, the most informative task analyses in this section are the suffix generalisation task, and the determiner and suffix generalisation task.

Most arbitrary mapping: vocabulary

Translation recognition task

As already mentioned in the previous subsection, this task assessed participants' knowledge of the trained words and their meanings. As a reminder, half of the trials contained the match (correct) English word-novel word pairing translations, half contained mismatch (incorrect) pairings. Participants' mean accuracy was .836 for correct trials, .855 for incorrect trials. The average A' discrimination of the immediate group was .899. Numerically, this indicates that participants tested immediately had a high level of accuracy on this task and a good level of discrimination between correct and incorrect trials.

Recall (stems)

As already mentioned in the previous subsection (and in Experiments 1-3), data from the recall task was analysed for stems and suffixes separately. As already mentioned in the previous subsection, stem recall was used to analyse participants' knowledge of the individual stem-picture mappings. For the immediate group, a within subjects t-test with the variable of cue type (picture only or picture and grapheme) was run. The dependent variable was stem recall accuracy (proportion correct out of total productions). There were no significant differences between the two cue types ($t(50)= 2.12$, $p = .039$). As shown in Figure 54, participants from the immediate group recalled more stems correctly when the grapheme cue was present, rather than when the picture only was present.

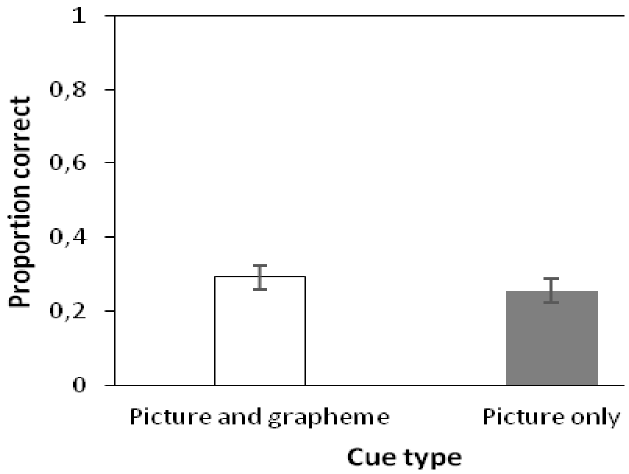


Figure 54. Stem recall accuracy for the immediate group.

Most systematic mapping: suffixes (grammar)

For the suffix tests, it was of interest whether the results would indicate that the suffix-gender regularities were available immediately.

Recall (suffixes)

As already mentioned in the previous subsection, the suffix recall was used to analyse participants' knowledge of the individual suffix-picture mappings. For the immediate group, a within subjects t-test with the variable of cue type (picture only or picture and grapheme) was run. The dependent variable was suffix recall accuracy (out of total productions). There were no significant differences between the two cue types ($t(50) = .29$, $p = .771$). As shown in Figure 55, participants tested immediately had similar levels of suffix recall accuracy regardless of cue type.

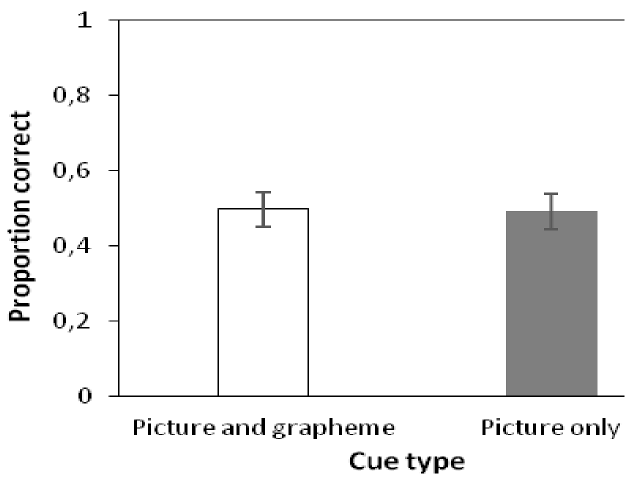


Figure 55. Suffix recall accuracy for the immediate group.

Suffix generalisation task

As already mentioned in the previous subsection, this task was used to assess participants' suffix-gender regularity extraction. It was of interest whether there would be evidence of suffix-gender regularity extraction in the immediate group. As a reminder, in this task we are interested in the pattern of endorsement of consistent compared to inconsistent items as indication of suffix-gender regularity extraction (as opposed to overall levels of performance).

For the immediate group, a within subjects t-test with the variable of consistency (consistent and inconsistent) was run. The dependent variable was endorsement. There were no significant differences in the endorsement of the two levels of consistency (t(50)= -.50, p = .621). As shown in Figure 56, participants from the immediate groups had a similar level of endorsement regardless of consistency, hence providing no evidence of suffix-gender regularity extraction immediately.

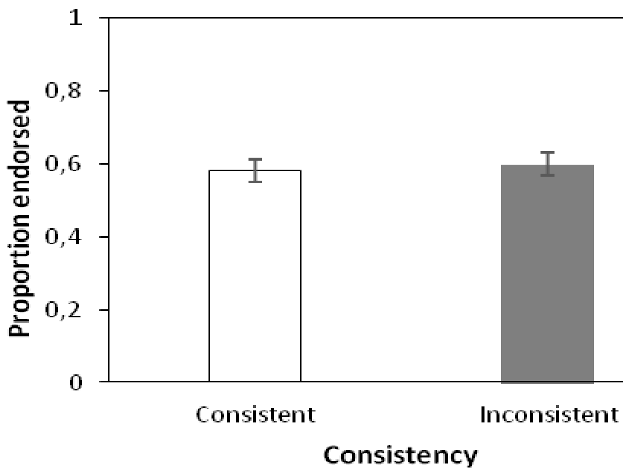


Figure 56. Suffix generalisation performance for the immediate group.

Next, as with Experiments 1-3, and as with the analysis of the delayed group in Experiment 4, one sample t-tests against chance (0.5) were run, separately for consistent and inconsistent items. Participants from the immediate group endorsed consistent items significantly above chance (t(50)= 2.52, p = .015). However, participants from the immediate group also endorsed inconsistent items significantly above chance (t(50)= 3.12, p = .003). This pattern is not in line with the pattern expected if there was suffix-gender regularity extraction. Hence, there was no evidence of suffix-gender regularity extraction immediately. However, given that participants were above chance on endorsing both consistent and inconsistent items, they may have been relying on the extraction of the determiner-gender regularity to complete the task. The results of the determiner and suffix generalisation task (presented in the next section) will confirm whether this is the case.

Overall, there was no evidence of suffix-gender regularity extraction immediately. Firstly, there were no significant differences in endorsing consistent and inconsistent items. Secondly, the results of one-sample t-tests against chance indicated that participants' pattern of endorsement was not in line with the pattern expected for suffix-gender regularity extraction. This indication of immediate performance is informative for our findings in the delayed groups.

As a reminder, we have not seen strong evidence of suffix-gender regularity in Experiments 1-3. In addition, as illustrated in the analysis of the delayed groups, there was also no strong evidence of suffix-gender regularity extraction in Experiment 4 (i.e. in this Experiment, outlined in the previous Results subsection), albeit with *some* indication of extraction in the wake group only. The findings from the immediate group suggest that suffix-gender regularities were not extracted immediately, but then lost with time, but were simply not available immediately.

Intermediate systematicity mapping: determiners (grammar)

Determiner selection task

As a reminder, this task was used to assess participants' memory for the trained determiner-picture mappings. For the immediate group, the average proportion of correctly selected determiner-picture mappings was .414. Hence, memory for individual determiner-picture mappings was low. The suggestion that memory for individual determiner-picture mappings was low is further supported by looking at absolute performance. The average accuracy count out of 16 items was 6.6 for the immediate group, indicating that memory for individual determiner-picture mappings was low. Nonetheless, according to the one sample t-test against chance (.25), performance was significantly above chance ($t(50)= 4.63$, $p < .001$). This suggests that although memory for individual determiner-picture mappings was low, the immediate group nonetheless had memory traces for the individual determiner-picture mappings from which the determiner-gender regularities could be extracted.

Analysis of incorrect responses on the determiner selection task

As with the analysis of both Experiment 3 and the delayed groups in Experiment 4, to investigate whether participants may have relied on the extracted determiner-gender regularity to complete this task, participants' errors were analysed. Errors reflecting correct determiner-gender mappings (proportion of correct gender determiners out of the incorrect but trained determiners) were compared against the chance level of 0.33.

Participants were not significantly above chance at selecting the correct determiner-gender mapping when making an error ($t(50)= -1.26$, $p = .212$), the mean proportion was .31. This suggests that not only participants' individual item memory for determiner-picture mappings was low, but participants also did not show any evidence of relying on the determiner-picture gender *regularity* to complete this task. (This is also in line with the findings for the two delayed groups in Experiment 4).

Determiner and suffix generalisation task

As a reminder, this task was used as an additional test of the extraction of the determiner-gender mappings. It was of interest whether there would be extraction of determiner-gender regularities immediately.

For the immediate group, a within subjects t-test with the variable of consistency (items consistent versus inconsistent with the trained regularities) was run. The dependent variable was proportion endorsed. There was a significant difference between the two levels of consistency ($t(50)= 2.42$, $p = .019$). As shown in Figure 57, in the immediate group, there was higher endorsement of consistent items compared to inconsistent, providing evidence for determiner-gender regularity extraction.

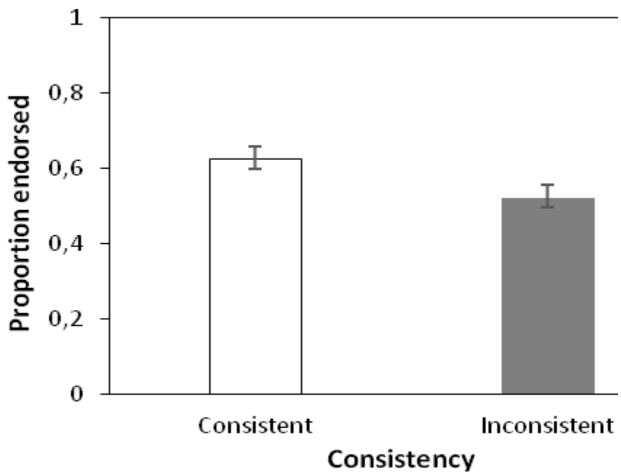


Figure 57. Determiner and suffix generalisation task performance for the immediate group.

As with both Experiments 1-3, and the analysis of the delayed groups in Experiment 4, performance was additionally compared against the chance level of 0.5. Participants in the immediate group were significantly above chance on endorsing the consistent items ($t(50)= 4.54$, $p < .001$), and at chance at endorsing the inconsistent items ($t(50)= .67$, $p = .509$). This pattern of endorsement is in line with that expected for determiner-gender regularity. The results of the one sample t-tests against chance are also in line with the significant differences between the two levels of consistency (presented above) suggesting that the determiner-gender regularity was extracted immediately.

This evidence for determiner-gender regularity extraction provides two suggestions. Firstly, it suggests that participants in the immediate group may have been relying on the determiner-gender regularities to complete the suffix generalisation task, as suggested earlier. (As a reminder, participants in the immediate group endorsed both consistent and inconsistent items above chance on the suffix generalisation task presented earlier in this subsection).

Secondly, the evidence of determiner-gender regularity extraction on the determiner and suffix generalisation task is not in line with the analysis of errors in the determiner selection task. This difference may be as a result of the differences in the two tasks. As a reminder, in the determiner selection task, participants heard a trained word and saw a trained picture on screen. Their task was to select the determiner which corresponded to the word-picture mapping. Therefore, participants could have completed the determiner selection task by relying on their memory of the individual mappings, more specifically the co-occurrence of the individual determiner with the individual word, or the co-occurrence of the individual determiner with the individual picture. It was not possible to rely on memory of the *individual* trained mappings to complete the generalisation task. Rather, participants would have had to rely on the extraction of the determiner-gender regularities to complete the task.

Hence, this difference may reflect that participants in the immediate group had relatively poor memory for the *individual* determiner-picture mappings, but extracted the determiner-gender *regularity*. This would be in line with previous findings that forgetting of individual elements may be needed to facilitate generalisation (e.g. Vlach, Ankowski, &Sandhofer, 2012).

Overall Summary

It was of interest, whether the extraction of suffix-gender regularities and determiner-gender regularities would be available immediately, but lost following a consolidation delay.

For determiners, there was some evidence of determiner-gender regularity extraction in the immediate group, based on the findings from the determiner and suffix generalisation task. For suffixes, there was no *strong* evidence of extraction immediately.

Compiled with the findings from the two delayed groups, this suggests that determiner-gender regularities were available immediately, but lost following a consolidation delay. On the basis of one sample t-tests against chance, there was *some* evidence of determiner-gender regularity extraction being forgotten less in the wake group than in the sleep group.

The suffix-gender regularities were not available immediately, as evidenced by *both* the suffix generalisation task, *and* the one sample t-tests against chance. On the basis of one-sample t-tests against chance, there was *some* evidence that suffix-gender regularities begun to emerge with wake-dependent consolidation only (and not with sleep-dependent consolidation). Presumably, in order for the suffix-gender regularities to emerge *immediately*, a different language input structure might be required.

4.4.3 Discussion

Two separate questions were of interest in Experiment 4. Firstly, it was of interest whether any group differences (wake compared to sleep) emerged as a result of additional training. Secondly, it was of interest whether determiner-gender regularities and suffix-gender regularities were available immediately, but lost following a consolidation delay. The discussion is structured so as to address these two questions separately.

Delayed group differences

The interesting difference between Experiments 3 and 4 was in the amount of training that the participants received. Both of the Experiments had the same systematicity of mappings (4 determiners, 2 suffixes).

In terms of vocabulary, there was no evidence that additional training benefited the relatively higher recognition memory, or individual stem recall. There was a trend for increased suffix recall with increased training. Despite this trend, there was still no evidence of suffix-gender regularity extraction, and no evidence that increased training benefited suffix-gender regularity extraction. Additionally, memory for the individual determiner-picture mappings was low, and there was no evidence of determiner-gender regularity extraction.

Overall, these findings may suggest that additional training is not sufficient to compensate for the effect that swapping the systematicity of the mappings had on consolidation of the individual form-picture mappings and the extraction of form-gender regularities. Moreover, the findings potentially point to a difference in the way that determiner-gender regularities and suffix-gender regularities are processed for extraction, given that despite suffixes being fully systematic there was still no strong evidence of suffix-gender regularity extraction.

To elaborate on this point, the concepts of systematicity and frequency that we have been addressing may be related to the concepts of cue reliability and cue availability from the Competition model (e.g. Bates & MacWhinney, 1989), as already suggested in the introduction of this Chapter. The Competition model suggests that cues with higher reliability and higher availability are more likely to be extracted. From our findings, this appears to hold true for the extraction of determiner-gender regularities following a *consolidation* delay. More specifically, (regardless of sleep, and regardless of the training amount), after a delay determiner-gender regularities are extracted only when they are a more reliable and a more available cue within the input (to use the terms of the Competition model). The same does not appear to be true for suffixes, potentially pointing to a difference in the way determiner-gender regularities and suffix-gender regularities are processed for extraction. Potentially the presence of the determiner-gender regularities was inhibiting the extraction of the suffix-gender regularities. This will be addressed further in Chapter 5.

It is important to also mention that in Experiment 4 there was no statistical evidence for a benefit of sleep for the acquisition of individual stems (unlike Experiments 1 and 2). There was also no statistical evidence for a benefit of sleep for the acquisition of individual suffixes, as measured by suffix recall (unlike Experiments 1-3).

Immediate extraction of regularities

It was of interest whether the extraction of suffix-gender regularities and determiner-gender regularities would be available immediately, but lost following a consolidation delay. There was evidence to suggest that determiner-gender regularities were available immediately, but lost following a consolidation delay. There was *some* indication that wake-dependent consolidation, but not sleep-dependent consolidation, preserved some of the determiner-gender regularity extraction from being forgotten. In addition, for the immediate group, there was evidence of determiner-gender regularity extraction on the generalisation task, but not in the analysis of errors on the determiner selection task. This might be because participants could complete the determiner selection task by relying on memory of the individual and trained determiner-picture mappings. This was not possible on the generalisation task, where participants had to rely on their memory of the extracted determiner-gender regularities. Hence, participants in the immediate group may have had relatively poor individual determiner-picture mapping memory, but have extracted the determiner-gender regularity. This suggestion is in line with evidence suggesting that individual elements may inhibit generalisation, and hence forgetting of individual items may facilitate generalisation (e.g. Vlach, Ankowski, &Sandhofer, 2012; Werchan & Gomez, 2014).

The suffix-gender regularities were not extracted immediately. There was *some* indication that suffix-gender regularities *started* to emerge with wake-dependent consolidation, but not with sleep-dependent consolidation. It could be that suffix-gender regularity was inhibited by the presence of determiners in the input, which *also* cued the same gender mapping. Indeed, there was *some* indication that in the suffix generalisation task participants from the immediate group may have been relying on the extracted determiner-gender regularities. This is because even for items where suffixes were inconsistent with the trained regularities, the determiner-gender regularity remained consistent (as already mentioned previously). This tentatively suggests that the participants from the immediate group *may* have been preferentially processing determiners over suffixes in this task. Consequently, it is unclear whether had the determiners not been present in the input, participants might have adjusted their performance to account for the suffix-gender regularities.

Finally, on the basis of the Competition model (e.g. Bates & MacWhinney, 1989), as already mentioned in the previous section of the discussion, it could be suggested that the mappings which are more reliable and more available (i.e. more systematic and more frequent) in the input, are more likely to be extracted, regardless of sleep. It is interesting to mention that in the immediate group, we saw evidence of extraction of the relatively *less* systematic (and less frequent) determiner-gender regularity, but no evidence of the relatively *more* systematic (and more frequent) suffix-gender regularity. This is not in line with the suggestion that the more systematic and more frequent mappings are more likely to be extracted. This discrepancy potentially suggests that the presence of the determiners in the input was inhibiting the extraction of the suffix-gender regularities, as already suggested. In addition, this discrepancy potentially points to key differences in the way that determiners and suffixes are processed (as already mentioned), and suggests that key language input factors *other* than systematicity and frequency (i.e. other than cue reliability and cue availability) may determine whether regularities are extracted. This suggestion will be discussed further in Chapter 5.

4.5 Chapter Summary and General Discussion

The primary aim of Experiments 3 and 4 was to investigate the effect of the systematicity level (and hence frequency level due to the nature of the stimuli) on the extraction of regularities within systematic mappings. To summarise, regardless of sleep, the determiner-gender regularity was extracted in Experiments 1 and 2, when the mapping was more systematic. There was no *strong* evidence of determiner-gender regularity extraction in the delayed groups in Experiments 3 and 4, even despite additional training, when the mapping was less systematic. There was no *strong* evidence of suffix-gender regularity extraction neither in Experiments 1 and 2, when the suffix mapping was less systematic, nor in Experiments 3 and 4, when the suffix mapping was more systematic, even despite additional training. For the arbitrary measures, there was no strong evidence of a benefit of sleep for individual stem recall, unlike Experiments 1 and 2. Experiment 3 (but not Experiment 4) provided some evidence for the benefit of sleep for individual suffix recall.

Secondly, it was of interest whether the determiner-gender regularities and the suffix-gender regularities were available immediately, but lost following a consolidation delay. There was evidence to suggest that determiner-gender regularities were available immediately, but lost following a consolidation delay, with *some* preservation following wake-dependent consolidation, but not sleep-dependent consolidation. There was also evidence to suggest that suffix-gender regularities were *not* available immediately, with *some* indication that the suffix-gender regularity extraction *started* to emerge following wake-dependent consolidation, but not sleep-dependent consolidation.

Extraction of regularities regardless of sleep

It was expected that making the suffixes fully systematic would boost suffix-gender regularity extraction. Yet, there was no strong evidence supporting this. This suggests that the degree of systematicity is not the crucial factor that affects extraction of the suffix-gender regularities. Additionally, even increasing training did not boost suffix-gender regularity extraction. It could be that the level of training does not affect suffix-gender regularity extraction. Alternatively, the efficacy of the training task or format could be challenged. It is also possible that participants simply reach a plateau, and even more training (regardless of the format) would not boost acquisition.

However, these conclusions are complicated by the increased complexity (as a result of decreased systematicity) of the determiner-gender regularity mapping (i.e. having 2 determiners per gender is a more complex mapping than having only 1 determiner per gender). The increased complexity of the determiners may be masking any learning of the suffixes, given that both the determiners and the suffixes cue the *same* gender mapping.

To elaborate on this suggestion, there are for instance examples of sentences where the tense can be derived from *both* a temporal adverbial (e.g. "yesterday"), and the grammatical morphology (e.g. the regular verb ending "ed" signifies past tense), such as in the sentence "*Yesterday* I walk*ed*.". It has been shown that second language learners typically learn to derive the tense based on the temporal adverbial cue alone, and require a much longer time period to then acquire the (relatively redundant) grammatical morphology cue (e.g. Bardovi-Harlig 2000). This might be because on the basis of the native English participant's experience with their first language, they might have come to interpret grammatical *ending* cues as relatively less salient (e.g. Ellis, 2006). Given that our participants were native English speakers (as already mentioned in the Method of each Chapter in this thesis), they *may* have interpreted suffixes (as the grammatical *ending* cue) as less salient than determiners, leading to the learning of suffixes being masked by the determiners. Indeed, there was evidence from the immediate group in Experiment 4 that the determiner-gender regularities *were* extracted immediately, despite determiners being the less systematic mapping, but that suffix-gender regularities were *not* extracted immediately, despite suffixes being the more systematic mapping.

Moreover, the lack of immediate suffix-gender regularity extraction in Experiment 4 was again in contrast with the existing literature, which suggests that participants are able to extract regularities from sequences (e.g. Gomez et al., 2006; Saffran, Aslin & Newport, 1996; St Clair, Monaghan & Ramscar, 2009). Most of the studies above have focused on one domain only (phonology, i.e. strings of syllables), and not on cross-domain mappings, which could be an additional dimension of difficulty. Additionally, Van den Bos et al. (2012) have shown that fully systematic mappings are more likely to be acquired than probabilistic (i.e. less systematic) mappings. Yet, despite being fully systematic, suffix-gender regularities were still not extracted, even with the increased level of training provided in Experiment 4.

This contrast in findings raises the question of which is the crucial factor for extraction of regularities that wasn't met in our Experiments, that was met in previous studies which *have* found regularity extraction (e.g. Gomez et al., 2006). A possible factor is exemplar variability, as previous experiments (e.g. Gomez, 2002; Tamminen, Davis & Rastle, 2015) have found that increasing the number of exemplars which are consistent with the regularity to be extracted, boosts the extraction of that regularity. A second factor is the redundancy of multiple cues. It may be that having more than one cue for the same category (such as in the case of determiners and suffixes both cueing the same gender mapping), makes one of those cues redundant. This makes it unlikely that both cues would be extracted (e.g. Ellis, 2002). It is possible that the extraction of suffix-gender regularities was blocked by the presence of determiner-gender regularities. These two factors: exemplar variability and multiple cue redundancy will be investigated in the next Chapter.

Role of sleep in consolidation of arbitrary versus systematic mappings

As we have already discussed in the previous Chapters, the CLS model (McClelland et al., 1995) when applied to our Experiments, suggests that the benefit of sleep for consolidation comes from its role in aiding the hippocampal-neocortex information dialogue. Therefore, information which is more hippocampally-dependent (i.e. arbitrary more than systematic), should be more benefited by sleep, as this information would be more reliant on the hippocampal-neocortical dialogue for long term consolidation. Accordingly, it was expected that sleep should benefit the consolidation of individual arbitrary mappings (stems and suffixes) but not the extraction of regularities.

Indeed, individual suffix recall, but not extraction of the suffix-gender regularity, was benefited by sleep in Experiments 3, as expected based on the CLS, but not in Experiment 4. Additionally, there was no benefit of sleep for the extraction of determiner-gender regularities or for the extraction of suffix-gender regularities in both Experiments 3 and 4, as expected based on the CLS. Given that at the level of individual item learning, determiner-picture mappings were learned above chance, this suggests that a sufficient memory trace was acquired, based on which sleep could have boosted the extraction of the determiner-gender regularity from that memory trace. Yet, this was not the case, suggesting that sleep really does not aid determiner-gender regularity extraction. Hence, Experiments 3 and 4 have provided further evidence, that consistent with CLS predictions, while sleep might benefit the consolidation of arbitrary information, it does not benefit the consolidation of systematic information. There was no indication that the initial level of training or the degree of systematicity are two factors which mediate the role of sleep in extraction.

Although there was a numerical advantage for stem memory following sleep, the sleep benefit for arbitrary stems did not reach significance. This lack of sleep benefit for arbitrary stems is also accompanied by significantly higher stem recall in Experiment 3 compared to Experiment 1 (when the sleep benefit for arbitrary stems was observed). On the basis of this higher stem recall, it might be argued that making the determiners less systematic, brought more attention to the stems. This increased attention to the arbitrary stem mapping might have resulted in less need for sleep-related consolidation. Having said this, stem recall was still very low.

An alternative explanation, is that increased resources may have been devoted to learning and consolidating stems regardless of sleep, as phonological forms at the level of the individual word were simplified (due to decreased number of suffixes). This idea that more attention was given to the phonology at the beginning of the word is supported by the finding that the grapheme cue resulted in significantly higher recall in Experiment 3, than in Experiment 1. Therefore, given the higher stem recall in Experiment 3 compared to Experiment 1, the stem mappings in Experiment 3 *may* have been better integrated in the neocortical systems. Hence, an additional benefit of sleep may have been needed less in Experiment 3, compared to Experiment 1 - for consolidation.

Systematicity and frequency

One of the aims of Experiments 3 and 4 was to address the factors of systematicity with frequency. Specifically, because of the nature of our stimuli, the more systematic mapping was also the more frequent mapping in the input. Consequently, we could not definitively say that in Experiments 1 and 2 determiner-gender regularities were acquired purely because they were the more *systematic* mapping (as opposed to the more *frequent* mapping in the input). In Experiments 3 and 4, each suffix as a result of being more systematic was also seen more frequently in the input, than each determiner (i.e. as there was 1 suffix *per* gender, but 2 determiners *per* gender, each suffix-gender mapping was seen 8 times, i.e. more frequently, while each determiner-gender mapping was seen 4 times, i.e. less frequently).

Note, that the number of repetitions of each item was the same in Experiment 1 (and in Experiment 2) as in Experiment 3, making the results comparable (there were overall more repetitions as a result of additional training in Experiment 4). Suffix-gender regularities were not extracted in Experiments 3 and 4, despite the matched frequency of suffixes in Experiment 3 and 4 to determiners in Experiments 1 and 2. The more systematic and more frequent *suffix*-gender regularities were not extracted in Experiments 3 and 4, despite the more systematic and more frequent *determiner*-gender regularities being extracted in Experiments 1 and 2. This suggests that despite matched systematicity and frequency, determiners and suffixes are processed differently. Moreover, it suggests that the factors within the language input (such as degree of systematicity), as opposed to sleep, determine whether regularities are extracted or not. Hence the reasons why suffixes might not have been extracted will be addressed in the next Chapter (as already mentioned).

Conclusions

This chapter has not provided strong statistical evidence to confirm the findings from Experiments 1 and 2 that the consolidation of the individual arbitrary mapping is benefited by sleep. Nonetheless, consistently with the findings of Experiments 1 and 2, *and* the predictions on the basis of the CLS, extraction of regularities was not affected by sleep. Regardless of sleep, making the determiners less systematic inhibited their extraction and generalization following a consolidation delay, and was not compensated for by increased training. Making the suffixes more systematic still did not boost their extraction and generalization. It appears that the extraction of the suffix-gender regularities is influenced by crucial factors within the language input, other than the degree of systematicity. Two such factors could be cue redundancy, and also exemplar variability. Hence, these two factors will be investigated in the next Chapter.

Chapter 5: The redundancy of multiple cues

5.1 Introduction

The aim of Experiments 5 and 6 was to focus on the extraction of the suffix-gender regularity, in isolation of the determiners. Experiments 1-4 have already answered the question of determiner-gender regularity extraction. Specifically, as previously discussed in Chapter 4, findings from these Experiments suggest that following a consolidation delay, there is extraction of the determiner-gender regularity when they are fully systematic. There was evidence to suggest that despite the decreased systematicity, determiner-gender regularity extraction was available immediately, and then forgotten over the consolidation delay. There was *some* evidence to suggest that the wake-dependent consolidation process (but not the sleep-dependent consolidation) preserved at least some of the determiner-gender regularity extraction from being forgotten. Hence, the level of systematicity was the crucial factor which affected the level of determiner-gender regularity extraction following a consolidation *delay*.

Unlike determiners, there was no evidence of suffix-gender regularity extraction, regardless of their systematicity in Experiments 1-4. Even when the suffix-gender regularity was fully systematic, the extraction was still not available immediately. There was *some* indication that this extraction might have started to emerge following the wake-dependent consolidation process only. The findings potentially point to the extraction of the suffix-gender regularity being inhibited by the extraction of the determiner-gender regularity (particularly as evidenced by the immediate group's performance in Experiment 4). This is because both determiners and suffixes provided cues to the same category (i.e. gender), potentially making the suffix cue redundant. Hence, this Chapter primarily asks whether suffix-gender regularity extraction would be available if determiners are removed from the input, hence indicating that in Experiments 1-4 suffix-gender regularity was inhibited by determiner-gender regularity extraction.

Cue redundancy

The lack of suffix-gender regularity extraction in our Experiments is not consistent with the statistical learning literature, which provides evidence for extraction of statistical regularities. Crucially, the large majority of the studies in statistical learning literature have investigated the extraction of distributional regularities in the phonological form alone (e.g. Gomez, Bootzin & Nadel, 2006; Newport & Aslin, 2004, also see Romberg and Saffran (2010) for a review of statistical learning literature). This is a key difference from our studies, as we have provided semantic cues as well phonological (and distributional).

In line with both the multiple cue integration hypothesis (e.g. Allen & Christiansen, 1996), and a similar hypothesis applied cross-linguistically proposed by Monaghan, Christiansen and Chater (2007), the acquisition of grammatical categories is aided when multiple probabilistic cues (e.g. phonological, distributional and semantic) are combined, as opposed to encountered in isolation. Indeed, Brooks, Braine, Catalano, Brody and Sudhalter (1993) have found that phonological cues boosted the acquisition of arbitrary categories. The phonological cues were contained within noun labels referring to objects. Similarly, van den Bos, Christiansen and Misyak, (2012) have found that when non-adjacent dependencies (classically used in statistical learning literature) were probabilistic, as opposed to deterministic, acquisition was poorer. It was found that this could be compensated for by adding additional cues to the input, such as those contained within phonological regularities and visual referents. This provided further support for the view that multiple cues can aid the acquisition of regularities.

Yet, the lack of extraction across our Experiments suggests that this may not always be the case. Having multiple cues (i.e. phonological, distributional and semantic) in the input to category has not been sufficient in our Experiments for the extraction of both determiner-gender regularities and suffix-gender regularities to take place. This lack of extraction may be explained within the second language acquisition literature. This literature has focused not only on the regularities in the phonological domain (like the statistical learning literature), but rather encompassed the acquisition of cue-function mappings.

More specifically, the concept of redundancy as a consequence of multiple cues to the same function has received a lot of discussion within second language learning research and has been approached from different angles. A useful definition of cue redundancy is provided by DeKeyser (2005). By this definition, when a cue (such as the suffixes in our Experiments) is not necessary to understand the meaning of the word (and in this case the gender of the word), that cue is redundant. Hence, a discussion of redundancy is relevant for this Chapter, as it can be said that suffixes were a redundant cue in Experiments 1-4, given that the gender of the words could be derived from the determiners alone.

Van Patten's input processing model

Firstly, VanPatten has approached the concept of redundancy theoretically (rather than empirically) within the input processing (IP) model (and from a point of view of generative grammar). VanPatten (2004) has discussed what he terms "the preference for nonredundancy principle". This principle suggests that learners acquire redundant grammatical forms only after they have acquired the necessary (i.e. nonredundant) grammatical form. This preference for nonredundant forms arises out of the learner prioritizing the acquisition of individual item meaning (i.e. content form) over the acquisition of each individual grammatical form. In turn, this priority arises because of the learner's need to compensate for limited cognitive resources (e.g. Just & Carpenter, 1992). Hence, according to this model, the *more* crucial element to meaning, i.e. the "content form" should be processed first. The "grammatical forms" which are *less* crucial to individual meaning should be processed second. As a consequence, redundant grammatical cues may not be processed at all, in particular if they are also of low salience.

While the concept of meaning is not concretely defined within the IP model, its ideas may still be applied to our Experiments. In our Experiments, we have evidence from suffix recall, that suffixes *are* being processed. Yet, there is still no evidence that the suffix-gender gender mappings are being extracted, i.e. the more complex regularity. This might be explained by the presence of the more salient determiners as a cue to gender, making the less salient suffixes a redundant cue to gender, resulting in the learner's preference for determiners. (Reasons why determiners may have been interpreted as more "salient" than suffixes will be suggested in the next two sections of this Introduction). While relevant to this chapter, this view has remained largely a theoretical construct. The scarcity of empirical evidence directly for this model might potentially be explained by the difficulty of testing this model empirically.

The Competition model

The concept of redundancy is also related to aspects of the Competition model (e.g. Bates & MacWhinney, 1987; MacWhinney, 1987), which was introduced in Chapter 4. In Chapter 4, we considered the cue reliability and availability (i.e. systematicity and frequency) aspects of the Competition model (e.g. Bates & MacWhinney, 1987; MacWhinney, 1987). The Competition model suggests that mappings which are more reliable (i.e. more systematic) and more available (i.e. more frequent) in the input are more likely to be acquired. We found this to be true for determiners after a consolidation *delay*. As a reminder, after a consolidation delay we saw evidence of determiner-gender regularity extraction when determiners were the more systematic mapping in the input, but no strong evidence of determiner-gender regularity extraction when determiners were the less systematic mapping in the input. However, the same was not true for suffixes. As a reminder, we did not see extraction of the suffix-gender regularity, not even when suffixes were the more systematic mapping in the input. This suggests that language factors aside from systematicity and frequency (i.e. cue reliability and cue availability) may influence the extraction of the suffix-gender regularity.

Specifically, another key aspect to be considered aside from cue reliability and availability is the ordering of information in time. At the level of the sentence, in English specifically (but not in other languages, such as French and Italian), word order as a cue is assigned more strength for understanding the meaning, than morphological cues (such as subject-verb agreement) and semantic cues (e.g. MacWhinney, Bates & Kliegl, 1984). When competing cues are ordered differently in time, cues which occur earlier in the sentence, as a result of word order slow down the processing of cues appearing later in the sentence (e.g. MacWhinney & Bates, 1989). Although these findings are based on sentence (not individual item) comprehension, the same argument may apply to our Experiments. Participants may pay attention to the ordering of information in time. As determiners were encountered earlier in time than suffixes, due to being at the beginning of the word, the presence of the determiners in the input may have inhibited suffix-gender regularity extraction.

Another principle that applies is that of perceivability, i.e. how easy a cue is to identify. Perceivability has been shown to interact with cue validity. For example, in a study by MacWhinney et al., (1985) in Hungarian, although the accusative case marking is a cue with high validity, in some versions (i.e. when the case marking is less detectable phonologically on the basis of what phoneme it follows), it is assigned lower strength. That is because in some phonological examples, depending on which sound the marker follows (consonant or vowel), it is more difficult to identify, i.e. has lower perceivability. More perceivable cues are acquired earlier. This idea is relevant to our Experiments, given that determiners (unlike suffixes) are not only at the beginning of the word, but are also visually separate from the word. This is important, given that (as already mentioned in the Method in Chapter 2), our participants not only heard the words, but also saw their written word forms.

Consequently, the presence of determiners as a cue to gender on the basis of their greater perceivability and earlier ordering in time (due to being at the beginning of the word) may have made suffixes redundant as a cue to gender. Hence, removing determiners from the input may lead to the extraction of the suffix-gender regularities. To compare the Competition model (e.g. MacWhinney & Bates, 1989) to Van Patten's (e.g. 2004) IP model, the cue which has lower cue strength (and lower simple reliability and simple availability) within the Competition model, within Van Patten's IP model would be the redundant cue. The redundant cue (with lower cue strength) would only be processed, after the nonredundant cue (with higher cue strength) has been processed. Hence, removing determiners from the input may be *one* way to boost suffix-gender regularity extraction, by making suffixes a stronger cue (within the context of the Competition model), and a nonredundant cue (within the context of Van Patten's (e.g. 2004) IP model).

Preference for determiners over suffixes

There is reason to think that determiners may be selectively preferred over suffixes, and receive more attention than suffixes, due to being a more salient cue to gender. As already mentioned, determiners are at the beginning of the word (and also separate from the main word), whilst suffixes are at the end of the word (and also joined to the stem). There are several angles which suggest that attention is brought to beginning of a word, over its ending. Firstly, the processing heuristic (MacWhinney, 1999) suggests that learners pay more attention to items at the beginning of a word, than to endings. Secondly, as illustrated in Frigo and McDonald's (1998) study using grammatical markers that map on to gender, the factors of cue salience and position in the word may determine which of two cues is acquired first. For example in their Experiment 3, Frigo and McDonald (1998) had 8 conditions in a between-subjects design: there were word markers to gender within artificial words, either both in the beginning and end of the word (multiple cues resulting in redundancy), ending only, beginning only and no marking at all. Additionally, for each of those variabilities, each marker could either be highly salient (i.e. the marker was a whole syllable) or less salient (the marker was only a single phoneme).

To elaborate, there were 20 nonwords in the study, 10 per category (6 with regularities, 4 without). While we did not manipulate salience in our Experiments, in their study salience was manipulated by altering the number of syllables. In category 1, highly salient beginning regularities were "wan", and low salience regularities were "w". Salient ending regularities were "glot" and low salience regularities were "-t". The redundant condition included both the beginning and ending regularity, e.g. "*wan*ersum*glot*". Similarly to our Experiments, another set of generalization items was created based on the same regularities.

It was found that while all conditions resulted in above chance acquisition, generalization ability varied depending on the condition. All of the highly salient groups (regardless of the gender marker location) showed above chance generalization to new items. For the lower salience marker conditions, generalization performance was only above chance if a marker was present at the beginning of the word (i.e. both alone and in the redundant condition together with an ending marker). In the condition where the marker was only at the end, and was of low salience, generalization performance was not above chance. Moreover, participants relied on the beginning marker regardless of salience, but on the end marker only when it was more salient. Moreover, when during generalization participants were tested on words with only beginning markers and only end markers, performance was better for words with beginning markers. These findings suggest that learners process and generalize grammatical gender information at the beginning of the word, before they process and generalize grammatical gender information at the end of the word. Hence, cues to gender at the end of the word, such as our suffixes, may be treated as redundant, compared to cues at the beginning of the word, such as our determiners, at the very initial stage of language acquisition (such as investigated in our Experiments).

In addition, based on corpus analyses Monaghan, Christiansen and Fitneva (2011) have suggested that English speakers typically process information at the beginning of the word for the individual, arbitrary meaning. They process information at the ending of the word for the systematicity which maps on to the grammatical category. Monaghan et al.'s (2011) findings can be integrated into the models discussed above with relevance to our Experiments. For instance, Van Patten's (e.g. 2004) IP model suggests that learners prioritize acquiring individual meanings, over grammatical forms that map onto categories, out of the need to most efficiently assign limited cognitive processing resources.

Moreover, according to Rescorla and Wagner's (1972) model, the assignment of selective attention to one cue over another, may depend on its outcome importance. Naturally, the higher an outcome importance is, the more likely that cue is to be acquired. Hence, given that in our Experiments, we are looking at the very *initial* stage of language acquisition, at this stage the acquisition of individual word meanings may be prioritized over the acquisition of grammatical forms (as suggested by VanPatten's IP model (e.g. 2004)). In other words, the acquisition of individual, arbitrary word meanings, may have higher outcome importance (to use Rescorla and Wagner's (1972) terminology), than the acquisition of systematic regularities cueing to word gender (perhaps in particular because there was only the one category of gender in the input, but several different individual words to be discriminated between). Given that our participants are native English speakers, based on their experience with the rules of the English language, their attention may be drawn to the processing of the beginnings of words, over the endings of words, to accomplish this goal of individual meaning acquisition (on the basis of Monaghan et al.'s 2011 findings that meaning information is typically at the beginning of the word, while category information is typically at the ending of the word).

Hence, the priority to acquire individual meaning over category information, and the learner's expectation for this individual meaning information to be at the beginning of the word, may increase the determiners' salience over the suffixes' salience. This may lead to greater processing of determiners, over suffixes. It can be hypothesized, that the greater processing of determiners as the cue at the beginning of the word (and hence the more salient cue for the participants), may have inhibited the processing of suffixes as the cue at the ending of the word (and hence the less salient cue). This may be the reason why suffix-gender regularities were not extracted, in the presence of determiners cueing the same gender category.

In addition, it is crucial to remember that the goal of language-learning is to enable communication. This means that beyond just the structure of the language, speech, and in the case of our Experiments individual words, are also processed acoustically. Consequently, it is important to touch upon the importance of issues of salience within speech. As prosody can cue salience, determiners may have been interpreted as being more acoustically salient in comparison to the suffixes, on the basis of prosodic information.

For instance, the learner's attention is drawn to syllables which are stressed, or are pronounced with higher pitch or frequency, or are pronounced less quickly (e.g. Cutler & Fodor, 1979; Cutler & Norris, 1991). Although in our Experiments determiners were not stressed by the speaker for the Experimental stimuli recordings, the determiners may have potentially had higher frequency than suffixes. This may have made the determiners more acoustically salient than the suffixes. Moreover, the determiners (e.g. *"tib"*) also had a reduced, shorter vowel sound, in comparison to the suffixes (e.g. *"eem"*) which had the longer duration vowels. In addition, the determiners *may* have been acoustically parsed from the stem, unlike the suffixes (i.e. the speaker might have inserted a quick pause in between the determiner and the beginning of the stem). If this parsing (i.e. pause) occurred, this would have made the determiners more acoustically salient than suffixes. Hence, the determiners' greater acoustic salience in comparison to the suffixes, may have directed more attention to the determiners. An acoustic analysis would be required in order to consider this further.

To investigate whether determiners were inhibiting suffix-gender regularity extraction, Experiments 5 and 6 will have no determiners in the input in order to give participants the maximum opportunity to process suffixes for extraction. It is expected that there will be evidence of suffix-gender regularity extraction, regardless of sleep.

5.2. Experiment 5

Aims and predictions

The primary aim of Experiment 5 was to investigate the extraction of the suffix-gender regularities in isolation of the determiners. In Experiments 1-4, while the simple availability and reliability impacted the extraction of the determiner-gender regularities following a consolidation delay, there was still no evidence of suffix-gender extraction, despite suffixes being fully systematic (i.e. more reliable compared to determiners in predicting gender). Greater perceivability and earlier ordering in time may have made determiners more salient to participants, in comparison to suffixes. The presence of determiners as a cue to gender may have made suffixes redundant in predicting gender. As a result, the extraction of the suffix-gender regularities may have been inhibited by the presence of the determiners, due to competition between determiners and suffixes.

As a result of there being no determiners in the input, the learner's attention may be guided to suffixes, as the next strong cue to gender. Hence, Experiment 5 is an exploration of the effect of determiner removal on suffix-gender regularity extraction.

In terms of sleep-associated consolidation, consistent with the predictions of the Complementary Learning Systems model (McClelland et al., 1995), across Experiments 1-4 we found *some* evidence that sleep benefited the consolidation of the arbitrary mapping (although this was not *always* consistent), but not regularity extraction. Consistently with this, it is expected that in Experiment 5 sleep will benefit the consolidation of the arbitrary mapping, but not the extraction of regularities.

5.2.1 Method

The participant criteria, design, and procedure were in line with the previous Experiments. The key differences in method and a reminder of the general points for participant criteria and procedure are outlined below.

Participants

Fifty monolingual native English speakers from the University of York took part in this study. Out of these, one participant did not return for the second session, and therefore was excluded from analysis. There were 24 participants in the wake group, with a mean age of 19.88 ± 1.45 [SD]. Out of these, 4 were male, and 1 was left-handed. There were 25 participants in the sleep group, with a mean age of 19.56 ± 1.56 [SD]. Out of these, 8 were male, and 2 were left-handed.

Design

There were two groups (daytime wake and overnight sleep at home) as with Experiments 1 and 3.

Materials

Language tasks' stimuli

There were two key differences in the materials from the previous Experiments.

1. There were no determiners in the input, unlike Experiments 1-4. The consequence of this was that unlike Experiments 1-4, the determiner selection task was not included in the session two tasks.

2. There were altogether 2 suffixes, 1 per gender ("*eem*" and "*ool*").

Items were created by removing the determiners from the stem-suffix combinations used in Experiments 3 and 4. As with Experiments 1-4, overall there were 16 different words, 8 per gender. As with Experiments 3 and 4, there were eight stems per suffix, but this time without determiners. All spoken versions of the items without determiners were recorded by a male native English speaker.

Half of the 16 items ended in the suffix "*eem*", the remaining half of the items ended in the suffix "*ool*". Each suffix mapped on to the natural gender of a picture. As with the previous Experiments, this was counterbalanced across two versions of the language (i.e. for one version "*ool*" was masculine, *"eem"* was feminine, vice versa for the other language), e.g. "*darleem*" – "waitress", "*geechool*" – "soldier".

Experiment 5 kept the same level (of additional) training that was introduced in Experiment 4. This means that in the training task (i.e. the word-picture matching task), there were 384 trials in total, as with Experiment 4 (as opposed to 192 trials in total as with Experiments 1-3).

Generalisation items

The structure of the generalisation items was exactly the same as in the previous Experiments. In Experiment 5, there were 2 tasks, counterbalanced in order, with a total of 32 items (half consistent, half inconsistent). Having two generalisation tasks was in line with the design of Experiments 1-4, and also provided more statistical power. As determiners were removed from the input, both of the tasks tested the extraction of the suffix-gender regularity only.

Separating the two tasks allowed to compare whether participants' endorsement of consistent and inconsistent trials on this task changed across exposures. As the first generalisation task provided participants with additional exposure (beyond the training word-picture matching task in session one), it might have given participants an additional opportunity to detect firstly the phonology of the individual suffixes, and secondly the presence of two genders (represented by the male and female pictured characters) in the input.

Procedure

The procedure and order of the tasks was the same as in Experiments 1-4. As already mentioned in the Materials, the only difference in procedure from Experiments 1-4 was that there was no determiner selection task in session 2 (as the determiners were removed from the input).

5.2.2 Results

Language tasks

Training task (Session 1)

As with Experiments 1-4, the word-picture matching task was used to train participants on the new language mappings. Also as with Experiments 1-4, both accuracy at the end of training and correct RTs on target and noise trials separately were analysed. As a reminder, two versions of the language were used to counterbalance for suffix-gender mappings (e.g. "*ool*-male", "*eem*-female" in one version, vice versa in the second version), as with Experiments 1-4.

Accuracy analysis

As this was the first time the participants were exposed to the novel language, it was expected that participants from both groups and both language versions would reach similar levels of performance at the end of training, reflecting similar levels of familiarity with the language.

To analyse accuracy at the end of training, a two way between subjects ANOVA was run with the independent variables of group (sleep or wake) and language (version one or two). The dependent variable was accuracy (proportion of correct responses) at the final block (block 8) of the training task.

There was a significant main effect of group ($F(1, 47) = 4.10$, $p = .049$), no significant main effect of language version ($F(1, 47) = .82$, $p = .370$) and no significant interaction between group and language ($F(1, 47) = 2.29$, $p = .137$). As shown in Figure 58, participants in the wake group endorsed a significantly higher proportion of target trials as correct, than participants in the sleep group.

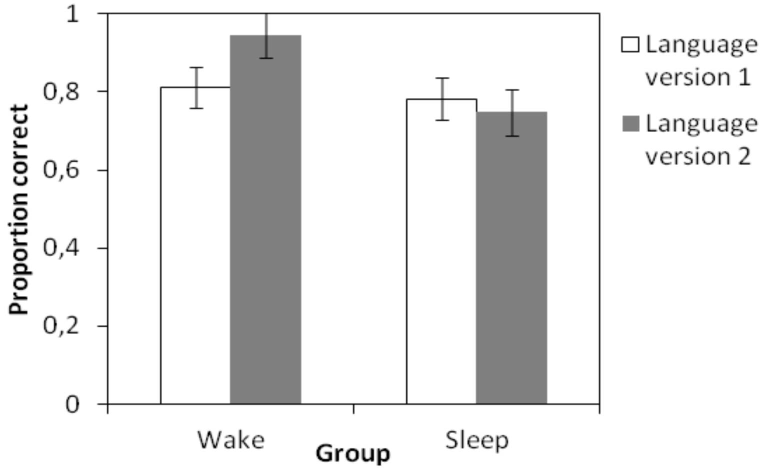


Figure 58. Accuracy at the end of training for the two groups and the two versions of the training language in Experiment 5.

Correct RTs analysis on the training task

As in the previous Experiments, an analysis of the RTs was additionally run to investigate the extent to which the difference between groups may be illustrative of learning. It was expected that participants' RTs should speed up with learning. To analyse this, instead of taking raw RTs, the RTs from the first block of training were subtracted from each of blocks 2 to 8. This was done for correct responses to target RTs (i.e. the hits), and separated for correct responses to noise RTs (i.e. the correct rejections).

A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, 8 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake/sleep).

Firstly, the dependent variable was the change in *target* RTs compared to block 1. One participant from the wake group did not have any correct responses for target RTs, and therefore could not be included in the analysis. Hence, the analysis was based on 23 participants in the wake group and 25 participants in the sleep group. There was a significant main effect of block ($F(1, 44) = 31.80$, $p < .001$), no significant main effect of group ($F(1, 44) = .16$, $p = .694$), and a significant interaction between block and group ($F(1, 44) = 2.22$, $p = .042$).

As shown in Figure 59, and as indicated by the block x group interaction, initially, at block 2, the sleep group's difference from block 1 in target RTs was greater than the wake group's. After this block, the wake group improved more across the blocks, than the sleep group, indicative of greater levels of learning in the wake group, than in the sleep group.

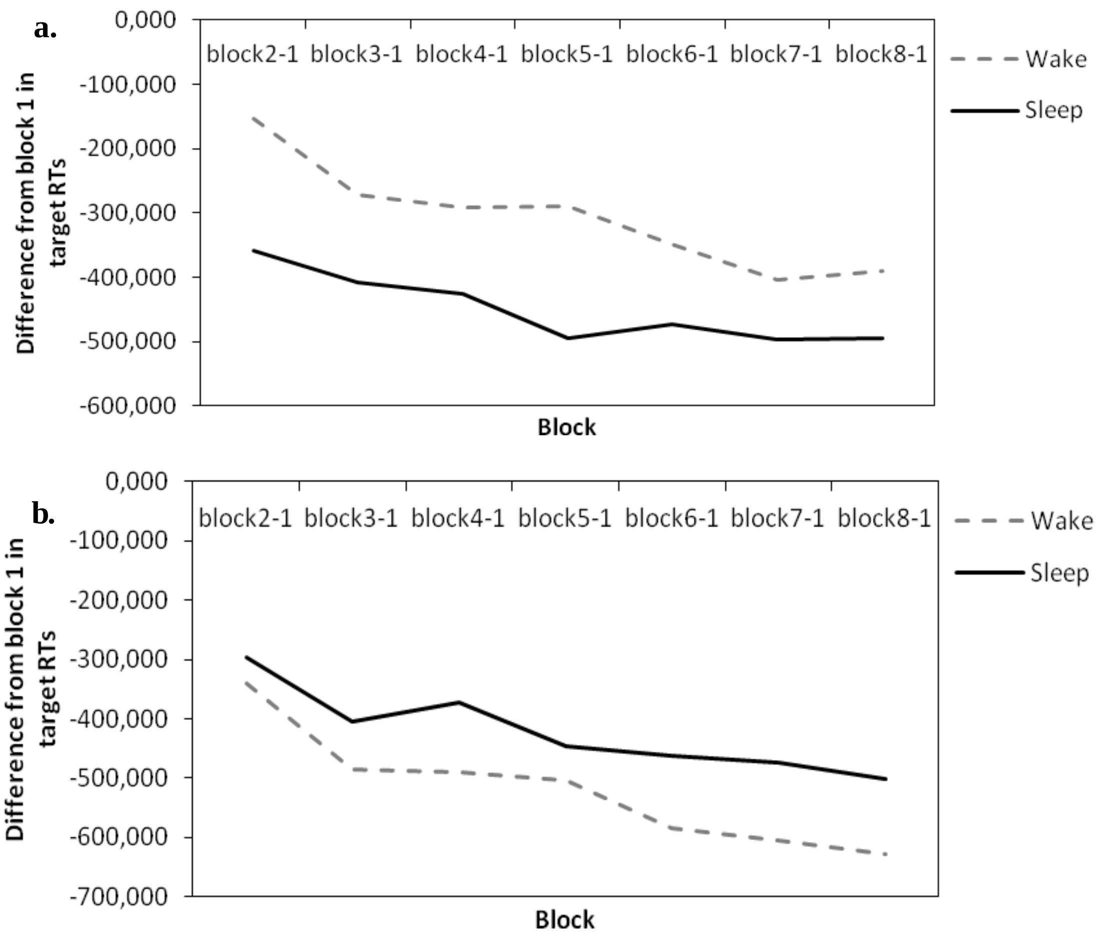


Figure 59. Difference in correct target RTs on the training task (Figure 59a. shows language version 1 RTs, Figure 59b. shows language version 2 RTs).

In addition, there was a marginally significant main effect of language version ($F(1, 44) = 3.86$, $p = .056$), a significant interaction between group and language version ($F(1, 44) = 6.55$, $p = .014$). There were no significant interactions between block and language version ($F(1, 44) = .67$, $p = .678$), or block, group and language version ($F(1, 44) = .10$, $p = .996$). As shown in Figure 59, and as indicated by the group x language version interaction, in comparison to the sleep group, the wake group had less improvement in target RTs on language version 1 (Figure 59a), but more improvement in target RTs on language version 2 (Figure 59b).

Next, to analyse performance on correct noise RTs (i.e. correct rejections), a mixed ANOVA was run, with the within subjects variable of block (difference in RTs between blocks 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, 8 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake/sleep).

The dependent variable was the change in noise RTs compared to block 1. One participant from the wake group did not have any correct responses for noise RTs, and therefore could not be included in the analysis. Hence, the analysis was based on 23 participants in the wake group and 25 participants in the sleep group. There was a significant main effect of block ($F(1, 44) = 16.83$, $p < .001$), no significant main effect of group ($F(1, 44) = .07$, $p = .797$), and no significant interaction between block and group ($F(1, 44) = 1.77$, $p = .106$). As shown in Figure 60, all participants showed improved noise RTs across the blocks, regardless of group.

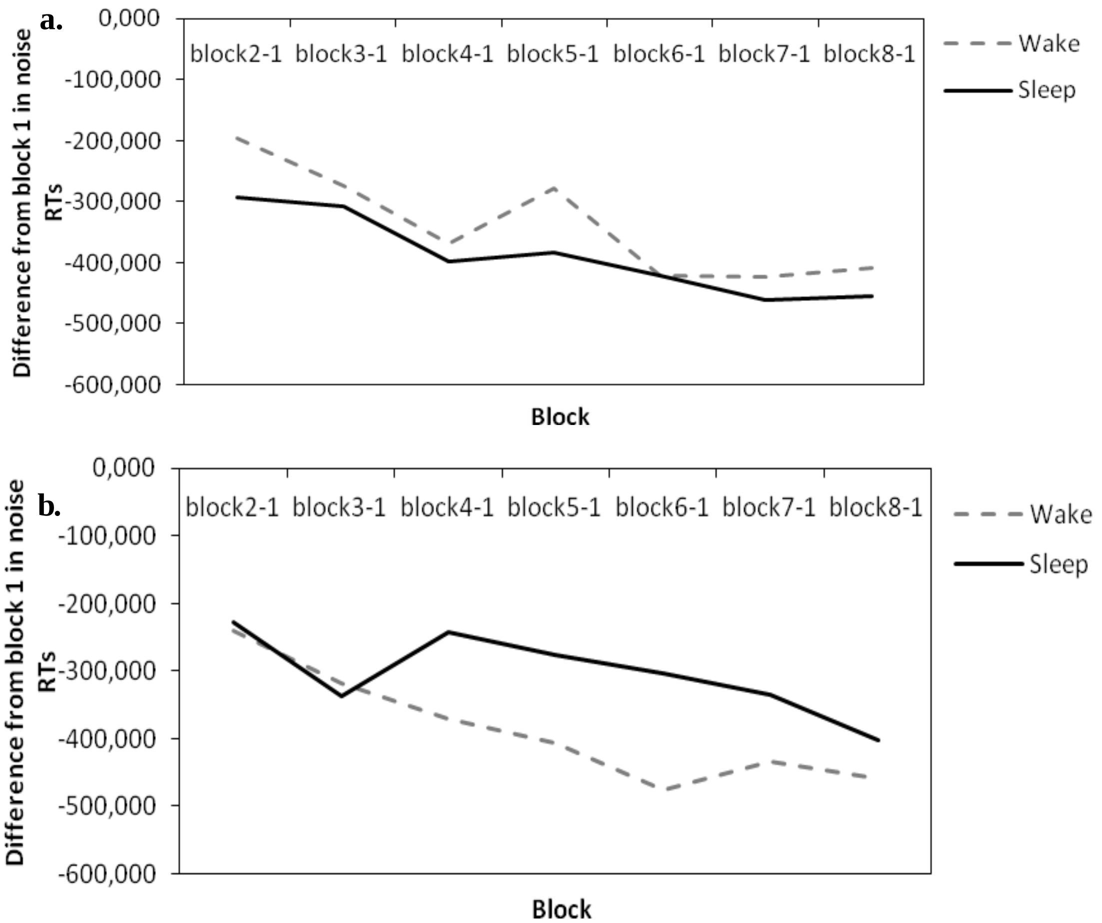


Figure 60. Difference from block 1 in noise RTs on the training task (Figure 60a. shows language version 1 RTs, Figure 60b. shows language version 2 RTs).

In addition, there was no significant main effect of language version ($F(1, 44) = .09$, $p = .767$), no significant interaction between group and language version ($F(1, 44) = 1.09$, $p = .301$). There were no significant interactions between block and language version ($F(1, 44) = 1.41$, $p = .211$), or block, group and language version ($F(1, 44) = 1.08$, $p = .378$).

To summarize, the wake group overall improved more than the sleep group across the blocks on the correct target RTs, indicative of greater learning in the wake group than in the sleep group. To account for the differences between groups at training, session 1 difference between block 8 and block 1 (i.e. the final block and the first block) for target RTs will be included in the session two tests as a covariate.

As the main interest of the investigation were group differences at the session two language tasks, session two analyses will be collapsed across language. Nonetheless, it is also important to mention that differences as a result of language version emerged on the target correct RTs analyses. Specifically, the wake group had less improvement in target RTs on language version 1, but more improvement in target RTs on language version 2, in comparison to the sleep group. This will be discussed further in the General Discussion, to see whether a pattern of differences according to language version emerged across all of the Experiments presented in this thesis.

Language tests (Session 2)

As a reminder, given the differences found on the training task, session 1 training was included as a covariate on all of the session two language tasks analyses. This measure was the difference between block 8 and block 1 on correct target RTs. Consequently, the analysis is based on 23 participants in the wake group and 25 participants in the sleep group (as 1 participant from the wake group did not have any correct responses for the target RTs during training).

Most arbitrary mapping: vocabulary

Translation recognition task

As in the previous Experiments, participants' memory of trained vocabulary was tested using a translation recognition task. The accuracy analysis revealed that participants in the wake group had an accuracy rate of 78.4% for match trials and 76.6% for mismatch trials. Participants in the sleep group had an accuracy rate of 80.5% for match trials and 86.0% for mismatch trials. This suggests that overall recognition memory was good.
A one-way between subjects ANCOVA was run with the independent variable of group (sleep or wake) and session 1 training (using the difference between block 8 and block 1

on correct target RTs) as the covariate. A’ discrimination was used as the dependent variable, in order to account for potential biases within participants’ responses to match and mismatch trials.

There was no significant main effect of group ($F(1, 44) = 1.63$, $p = .208$) and no significant main effect of session 1 training as the covariate ($F(1, 44) = .00$, $p = .966$). There was no significant interaction between group and session 1 training as the covariate ($F(1, 44) = 1.63$, $p = .208$).

Corrected A’ discrimination averages (taking session 1 training effects into account) were .852 for the wake group and .890 for the sleep group. This indicates that participants from both groups had a good rate of discrimination and identification between match and mismatch trials. Participants from both groups had similar performance, when session 1 training performance was controlled for, as shown in Figure 61.

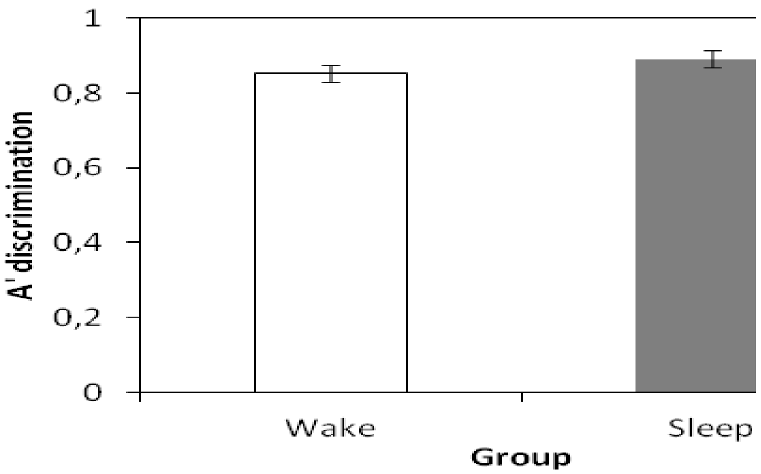


Figure 61. Performance on the translation recognition task in Experiment 5.

Recall (stems)

As in the previous Experiments, participants’ memory for trained vocabulary was also measured using a picture naming recall task. As in the previous Chapters, participants’ utterances of each word were separated into the stem and the suffix for analysis. This was done in order to separate the arbitrary mapping (stems) from the mapping with systematicity by design (suffixes).

Given that stems are the arbitrary mapping, it was expected that participants from the sleep group would have higher recall of stems than participants from the wake group, reflecting that sleep benefits the consolidation of arbitrary mappings. A mixed ANCOVA with a between subjects factor of group (wake/sleep) and within subjects factor of cue type (picture and grapheme/ picture only) was run with session 1 training as the covariate. Stem accuracy (proportion correct out of total productions) was the dependent variable. There was no significant main effect of group ($F(1, 44) = .34$, $p = .565$), and no significant main effect of cue type ($F(1, 44) = 2.09$, $p = .156$). There was no significant interaction between group and cue type ($F(1, 44) = 2.29$, $p = .595$).

As Figure 62 shows, participants from both groups recalled a similar proportion of stems correctly. Although numerically the sleep group's performance was overall slightly higher than the wake group's, there was no statistical evidence to suggest that sleep benefited stem recall, unlike in Experiments 1 and 2.

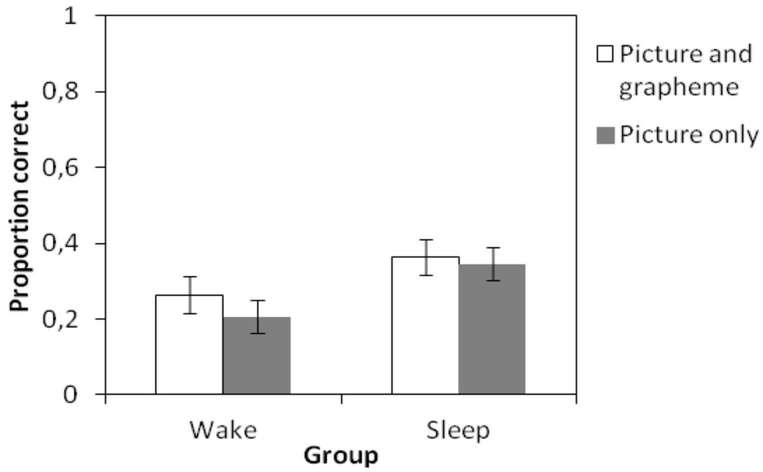


Figure 62. Stem accuracy on the recall task in Experiment 5.

Additionally, session 1 training as the covariate had no significant effect on performance ($F(1, 44) = .17$, $p = .679$). There was no significant interaction between session 1 training and group ($F(1, 44) = 1.73$, $p = .195$). There was no significant interaction between session 1 training covariate and cue type ($F(1, 44) = .45$, $p = .505$), and no significant three way interaction between session 1 training covariate, group and cue type ($F(1, 44) = .01$, $p = .931$). Hence, there was no evidence to suggest that session 1 training performance had an effect on stem recall, regardless of group.

To summarize, the findings from the vocabulary tests suggest that when the 12 hour consolidation delay included a period of sleep, both recognition memory and recall memory for the arbitrary mapping (stems) was *statistically* similar regardless of group, when session 1 training was controlled for. Hence, there was no evidence of a benefit of sleep for the consolidation of arbitrary mappings.

Systematic mapping: suffixes (grammar)

Recall (suffixes)

As with our previous Experiments, performance on the recall task was also used to compare the two groups at the systematic level (suffixes). It was expected that participants from the sleep group would have higher recall of suffixes at the individual level, than participants from the wake group, in line with the findings from Experiments 1-3 (although in Experiment 4 there was no statistical evidence for this sleep benefit). A mixed ANCOVA with a between subjects factor of group (wake/sleep) and within subjects factor of cue type (picture and grapheme/ picture only) was run with session 1 training as the covariate. Suffix accuracy (proportion correct out of total productions) was the dependent variable.

There was no significant main effect of group ($F(1, 44) = .38$, $p = .543$), no significant main effect of cue type ($F(1, 44) = .08$, $p = .775$) and no significant interaction between group and cue type ($F(1, 44) = .001$, $p = .971$). As Figure 63 shows, participants from both groups recalled a similar level of suffixes correctly, regardless of the cue provided (picture and grapheme or picture only) and regardless of the group. Although numerically overall suffix recall was slightly higher in the sleep group, than in the wake group, there was no statistical evidence to suggest that sleep benefited suffix recall.

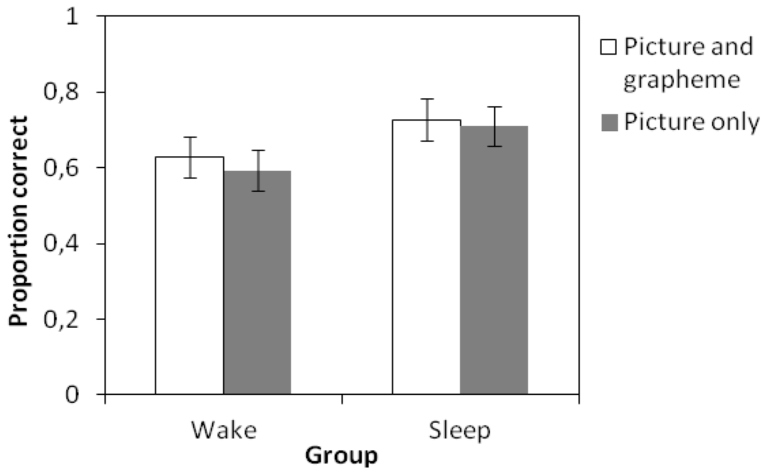


Figure 63. Suffix accuracy on the recall task in Experiment 5.

Additionally, session 1 training as the covariate had no significant effect on performance ($F(1, 44) = 2.09$, $p = .155$). There was no significant interaction between session 1 training and group ($F(1, 44) = 1.37$, $p = .248$). There was no significant interaction between session 1 training covariate and cue type ($F(1, 44) = .12$, $p = .744$), and no significant three way interaction between session 1 training covariate, group and cue type ($F(1, 46) = .04$, $p = .849$). Hence, there was no evidence to suggest that session 1 training performance had an effect on suffix recall, regardless of group. Experiment 5 did not show evidence for a benefit of sleep for suffix recall, unlike Experiments 1-3.

Suffix generalisation task

As in the previous Experiments, to test participants' ability to apply the suffix-gender regularities, they learnt during training to unseen items, we used a suffix generalisation task. In Experiments 1 – 4 there was no clear evidence of suffix-gender regularity extraction in the presence of determiners. It was of interest whether there would be evidence of suffix-gender regularity extraction, now that determiners were removed from the input, and the suffixes were fully systematic.

Two generalisation tasks were used in Experiment 5. As outlined in the Method, this was in line with the design of Experiments 1-4, and also provided more statistical power. As a reminder, given that there were no determiners present in the input, both of the tasks tested the extraction of the suffix-gender regularity. Separating the two tasks allowed to compare whether participants' performance on this task changed across trials. Given that the first generalisation task would provide additional exposure (beyond the training task in session one), it might provide participants with an additional chance to detect firstly the phonology of the suffixes, and secondly the presence of two genders in the input. To assess whether this might affect generalisation performance, task order was included in the mixed ANCOVA.

A mixed ANCOVA was run with the between subjects variable of group (sleep or wake) and the within subjects variable of consistency (items consistent or inconsistent with trained regularities), and task order as a within-subjects factor. Session 1 training was included as the covariate. Proportion endorsed (pressing 'match' to consistent and inconsistent items) was the dependent variable. As a reminder, in this task we are interested in the pattern of endorsement of consistent versus inconsistent items, as indication of generalisation ability.

There was no significant main effect of consistency ($F(1, 44) = 1.09$, $p = .302$), no significant main effect of group ($F(1, 44) = .66$, $p = .421$), and no significant interaction between group and consistency ($F(1, 44) = .23$, $p = .634$). There was a significant main effect of task order ($F(1, 44) = 5.72$, $p = .021$). There was no significant interaction between task order and group ($F(1, 44) = .17$, $p = .683$), and no significant interaction between task order and consistency ($F(1, 44) = .20$, $p = .656$). There was a marginally significant three-way interaction between task order, consistency and group ($F(1, 44) = 3.64$, $p = .063$).

As shown in Figure 64, there was no *strong* evidence of suffix-gender regularity extraction, regardless of group. Nonetheless, as reflected by the marginally significant interaction between task order, consistency and group, at task 1, participants from the wake group showed a pattern of endorsement that was more in line with suffix-gender regularity extraction, than participants from the sleep group. By task 2, the pattern of endorsement was further in line with extraction in the wake group, and even less in line with extraction in the sleep group.

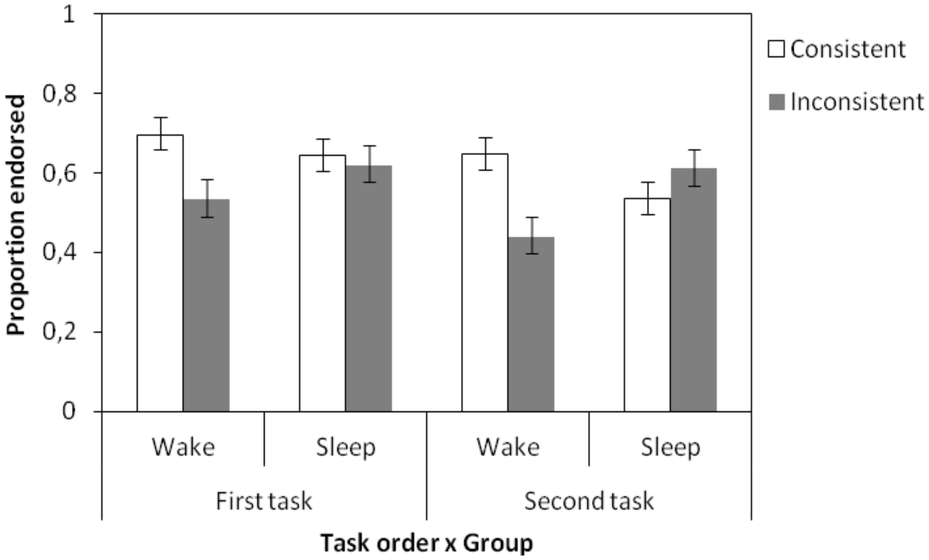


Figure 64. Suffix generalisation performance in Experiment 5.

Additionally, session 1 training as the covariate had no significant effect on performance ($F(1, 44) = .66$, $p = .421$) and there was no significant session 1 training covariate x group interaction ($F(1, 44) = .02$, $p = .885$). There was no significant interaction between session 1 training covariate and consistency ($F(1, 44) = .12$, $p = .748$), and no significant interaction between session 1 training covariate and task order ($F(1, 44) = 2.26$, $p = .140$). There were no significant three way interactions between task order, group and session 1 training ($F(1, 44) = .00$, $p = .994$) or between task order, consistency and session 1 training ($F(1, 44) = .07$, $p = .794$).

There was a significant four way interaction between consistency, task order, group and session 1 training ($F(1, 44) = 6.39$, $p = .015$). This interaction suggests that the wake group's pattern of endorsement being more in line with suffix-gender regularity extraction than the sleep group's, and in addition the wake group improving further by the second task, may be related to the wake's group greater improvement during the training task. Hence, it may be that the participants from the wake group (by chance, i.e. unintentionally) were overall better learners than participants from the sleep group.

As with Experiments 1-4, comparing performance against chance provides further evidence as to participants' suffix generalisation ability. As a reminder, if the suffix-gender regularity was extracted, participants should be significantly above chance (0.5) for endorsing consistent items, but also significantly below chance (0.5), or at least *at* chance for endorsing inconsistent items.

As shown in Table 14, one sample t-tests against chance (0.5) revealed that participants in the sleep group endorsed both consistent and inconsistent items significantly above chance on the first task, and only inconsistent items above chance on the second task. This pattern of endorsement in the sleep group was not in line with extraction. Participants in the wake group only endorsed consistent items above chance but were at chance for inconsistent items, in both the first and the second task. This pattern of endorsement in the wake group was in line with suffix-gender regularity extraction.

Overall, according to the one-sample t-tests against chance, the pattern of endorsement was in line with suffix-gender regularity extraction in the wake group, but not in the sleep group.

Table 14. Results of one sample t-tests against chance on the suffix generalisation task.

			t	df value	*p* - value	Mean proportion
Wake group	First task	Consistent	4.48	22	<.001	.70
		Inconsistent	.54	22	.598	.53
	Second task	Consistent	3.04	22	.006	.65
		Inconsistent	-1.19	22	.247	.44
Sleep group	First task	Consistent	4.13	24	<.001	.65
		Inconsistent	3.51	24	.002	.62
	Second task	Consistent	.96	24	.347	.54
		Inconsistent	2.86	24	.009	.61

Participants from the wake group (but not from the sleep group) may have started to extract the suffix-gender regularities. It appears more likely that this indication of extraction in the wake group is driven by the wake group's initially better acquisition at training, than as a result of wake-dependent consolidation processes. Indeed, the significant four-way interaction between consistency, task order, group and session 1 training may be in line with the suggestion that the wake group's extraction was driven by the better training performance. However, participants from the wake group had not yet *fully* extracted the suffix-gender regularity (given that inconsistent items were endorsed *at* chance, rather than significantly *below* chance). The suggestion that the wake group (but not the sleep group) may have started to extract the suffix-gender regularities is also tentatively supported by the marginally significant interaction between task order, consistency and group.

Although the task had novel stems and novel pictures, phonologically the suffixes were the same as at training. Moreover, while the specific gender occupation pictures were novel, the pattern of having two genders with human occupations was the same as at training. Consequently, for participants in the wake group who had begun to extract the regularities, the already familiar phonological suffixes and the presence of human occupation gender pictures in the input of the first task *may* have activated their memory traces for the suffix-gender regularities. This is may be a potential reason why performance improved by the second task.

For participants in the sleep group who had not yet begun to extract the suffix-gender regularities, the first task did not serve to improve performance. This might have been because there was no statistical regularity in the input in this task, i.e. half of the mappings were consistent, and the remaining half were inconsistent with the trained regularities. Consequently, the sleep group's performance decreased in the second task. It is important to bear in mind that this suggestion (that the wake group, but not the sleep group was starting to extract the regularities) is tentative, given that there was no significant group x consistency interaction in the generalisation task, which would have indicated that the wake group was better than the sleep group at generalisation.

Despite this, it is important to consider that these group differences may be starting to emerge – and with further exposure to the regularities might become more pronounced. However, it is equally possible (as already mentioned above), that these differences may *not* be down to the consolidation related processes of wake versus sleep. Rather, these differences may be emerging as a result of participants in the wake group being better learners in Experiment 5 (i.e. in this Experiment), due to individual differences (given the wake group's better learning during the training task, and the significant four-way interaction between consistency, task order, group and session 1 training, as suggested above).

Summary of language task results

Overall, the results have not provided statistical evidence that sleep is beneficial for the consolidation of arbitrary mappings at the individual level, as there was no benefit of sleep for stem recall (unlike Experiments 1 and 2), and no benefit of sleep on individual suffix recall (unlike Experiments 1-3). Consistent with Experiments 1-4, there was no evidence to suggest that sleep benefits the extraction of the suffix-gender regularities.

Regardless of sleep, despite suffix-gender regularity extraction not being "inhibited" by the presence of determiners, there was no strong evidence for suffix-gender regularity extraction. It is tentatively suggested that the pattern of endorsement was more in line with extraction in the wake group, than in the sleep group. As suggested above, this is likely due to the differences at training (with indication that the participants from the wake group performed better at training), rather than wake-dependent consolidation processes.

5.3 Discussion

The aim of Experiment 5 was to investigate whether the extraction of the suffix-gender regularity might have been "inhibited" by the presence of the determiners cueing the same gender mapping in our previous Experiments. In Experiment 5, there were no determiners present in the input, and suffixes were fully systematic. Despite this, there was no strong evidence for suffix-gender regularity extraction. It could be tentatively suggested that participants from the wake group (but not from the sleep group) began to extract the suffix-gender regularities. However, there was evidence to suggest that this is likely due to better performance at training, as opposed to as a result of the benefit of wake-dependent consolidation.

To elaborate, we found a four-way significant interaction between consistency, task order, group and session 1 training on the suffix generalisation task. This interaction indicated that on this suffix generalisation task, participants in the wake group showed performance that became even *more* in line with suffix-gender regularity extraction across additional exposures (hence the element of task order in the interaction), in opposition to the sleep group. In addition, (as indicated by the element of session 1 training performance in the interaction), this was related to the wake group also having higher performance at training. This overall suggests that the suffix-gender regularity extraction in the wake group may have been driven by their better learning, as opposed to being a result of wake-dependent consolidation processes.

This means that despite the predictions from Van Patten's IP model (e.g. 2004), and the Competition model (e.g. Bates & MacWhinney, 1987) discussed in the introduction, removing determiners from the input was not sufficient to guide the learner's attention to suffixes, as the next strongest cue to gender. This suggests that suffixes (even when fully systematic and not inhibited by the presence of determiners cueing the same gender mapping), do not have sufficient strength as a cue to gender. Hence, the salience of the suffixes to the learner needs to be increased. One possible way of doing this, is increasing exemplar variable (i.e. by increasing stem variability). Experiment 6 will seek to answer whether increasing stem variability will boost suffix-gender regularity extraction.

It is also of note, that there was no evidence for a benefit of sleep at the level of individual, arbitrary stems (unlike Experiments 1 and 2), or at the level of the individual suffixes (unlike Experiments 1-3).

5.4 Experiment 6

Importance of stem variability for learning and generalization

There is evidence suggesting that increasing exemplar variability (more so than increased frequency or repetition) aids the learning and generalisation of the regularity within language mappings. For instance, Gomez (2002) exposed adults (in their first of 2 studies) to 3-element artificial language strings, such as "pel wadim jic", which were comprised of a word order regularity. The regularity was contained within the non-adjacent dependencies. The first element predicted the last element, while the number of exemplars of the middle element varied. The degree of variability of the middle element was manipulated between 2, 6, 12 or 24 middle elements, while the overall frequency of the non-adjacent dependencies was kept constant across all conditions. Participants' detection of non-adjacent dependency regularities was assessed using a grammatical judgement test. Participants indicated whether the test strings (half of which were consistent with the trained regularities, half were inconsistent) were grammatically correct or not.

Gomez (2002) found that participants had stronger endorsement of consistent test strings (as opposed to inconsistent) items for the set with the variability of 24, compared to the set with the variability of 12. The same improvement in endorsement was not found when comparing the sets with lower variability (i.e. 2 and 6). Moreover, the improvement in endorsement of the set with the variability of 24 middle elements was not gradual, but radical. When the variability was lower (2, 6, and 12), participants did not show detection of the non-adjacent dependency regularities (as would have been evidenced through the larger endorsement rates of consistent as opposed to inconsistent test strings). This suggests that when the variability was lower, participants focused on adjacent dependencies, as opposed to non-adjacent dependencies. Findings suggest that greater variability within grammatical strings facilitates the learning of the word order regularity contained within nonadjacent dependencies in adults.

Similarly, Onnis, Christiansen, Chater and Gomez (2003) used the same stimuli as Gomez (2002). This time there were 5 possible set variabilities: 1, 2, 6, 12 and 24. Unlike

Gomez (2002)'s findings, Onnis et al. (2003) did not find that participants had stronger endorsement of consistent test strings (as opposed to inconsistent) items for the set with the variability of 24, compared to the set with the variability of 12. Instead, the effect was quadratic (i.e. U-shaped), with the greatest endorsement for set sizes of 1 and 24, and the lowest endorsement for set size of 6. Moreover, unlike the significant increase from set size 12 to 24 in Gomez (2002)'s study, Onnis et al. (2003) found the improvement in endorsement from set size 2 to size 24 to be gradual.

Similarly, Onnis, Monaghan, Christiansen and Chater (2004) also used the same strings. This time the variability of the middle unit was manipulated across 3 conditions: there were 1, 2 or 24 middle elements. In this study, the test strings contained novel items, as well as old items, half of which were grammatically consistent and the remaining half were inconsistent. As with Onnis et al. (2003)'s findings, the endorsement was U-shaped, with highest endorsement for sets of 1 and 24. This suggests that learning processes are similar for both when there is no variability (1 middle element), or variability is higher (i.e. 24 middle elements), but that the learning process is different when the variability is lower (i.e. 2 middle elements).

Onnis et al.'s (2004) Experiment 2 used different stimuli from their first Experiment. As previously, the stimuli were comprised of 3 parts, with variability of the middle part. However, this time the stimuli were words comprised of syllables, as opposed to strings. During the test, on each trial participants heard a "rule-word" and a "part-word", and had to indicate which of the 2 sounds was a "word". Participants endorsed "rule-words" over "part-words" in the higher variability condition (where the middle element varied between 24 syllables), but not in the lower variability condition variability (the middle element varied between 3 syllables). These findings suggest that higher variability boosts participants' ability to generalize the regularities contained within the non-adjacent dependencies, as evidenced by higher endorsement rates for "rule-words" over "part-words" in the higher variability condition.

The above findings from studies in artificial language have also been supported by findings from a natural language. Specifically, Eidsvag, Austad, Plante and Asbjornsen (2015) investigated adults' learning of Russian gender nouns. The nouns were composed of roots and gender-suffix regularities (e.g. in 1 possible combination, the root "kon" could be combined with either the masculine suffix "ya" or the masculine suffix "yem" to make up "konya" or "konyem"). It was found that learning was facilitated by increased variability of roots, as opposed to increased repetition during exposure, when the overall number of items was kept constant. Having 32 root words repeated once resulted in learning, but having 16 root words repeated twice did not. This was evidenced by only the participants who were exposed to increased variability of roots (as opposed to increased repetition) having greater endorsement of grammatically consistent as opposed to inconsistent items in the grammatical judgement test. The findings suggest that increased root variability facilitates the learning of grammatical gender-suffix regularities more than increased repetition. This is potentially because increasing the number of roots, makes the gender-suffix regularities more salient to the participant and hence easier to extract the regularity.

To summarize, the above studies (Eidsvag et al., 2015; Gomez, 2002; Onnis et al., 2003; 2004) provide evidence that increasing the variability within grammatical stimuli facilitates the learning of the grammatical regularity within that stimuli, both in an artificial language and in a real language. This is true for both adjacent dependencies, such as regularities contained within root-suffix words (as with Eidsvag et al., 2015), and for word order regularities contained within three-part grammatical strings within non-adjacent dependencies (as with Gomez, 2002; Onnis, 2002; 2003). The above studies have also provided evidence that immediately following learning, generalisation of regularities is greater when the exemplar variability was greater, as opposed to lower.

However, the unanswered question is whether the better performance seen during and immediately after learning as a result of greater variability, is still maintained after a consolidation delay. There is some evidence as to how increased stem variability may affect generalisation following a consolidation delay, in comparison to immediately after learning. Specifically, Tamminen et al.'s (2015) Experiment 2 investigated whether increasing the variability of stems to each paired affix boosts affix generalisation. Interpreted within CLS models (McClelland et al., 1995), novel affixes which have a larger number of stems paired with them, are likely to have been stored in pairing with a larger number of lexical representations, thus increasing the overlap of the affix as the common regularity across all of the items. Therefore, increasing the number of stems per affix increases the affix overlap in the neocortex, increasing the likelihood that the affix will be parsed from the stem and extracted as a common regularity.

In Tamminen et al.'s (2015) study, during training each of the 8 novel affixes was paired with either 8 stems or 2 stems to vary family size (e.g. the affix "*afe*" referred to a person, and examples of words ending in "*afe*" which could be included in a family included "sleep*afe*", "sail*afe*", "build*afe*", "teach*afe*", etc). The total number of times each affix was seen was the same across both conditions, such that in the condition with 2 stems per affix, the items were seen 4 times more frequently than in the condition with 8 stems per affix. Participants learned altogether 40 words – 32 in the more variable stems condition, 8 in the less variable condition. This allowed to parse episodic memory strength (which should be similar as affixes were seen the same number of items in both conditions) from affix extraction. Following exposure, participants were tested immediately with a free recall test (phonology only). Participants were tested on their recognition memory and ability to judge the accuracy of untrained words seen in sentences (sentence congruency) a week later following training.

In their study, immediately after training, participants recalled 48% of more variable items correctly, and 79% of the less variable items correctly, suggesting that either increasing variability or decreasing frequency of items decreases immediate individual memory for the *phonology* of items. Following a week of consolidation, recognition memory remained significantly higher in the less variable condition, potentially due to these items having been seen more frequently. Episodic memory for individual items was differently affected by consolidation than generalisation of regularities. In the sentence congruency task, testing generalisation, a congruency effect was seen for the higher variability condition, suggesting that generalisation is benefited by increased variability of stems per affix.

Experiment 6 in the current thesis would be an extension of Tamminen et al.'s (2015) study, as Tamminen et al., (2015) used already known stems (for which participants are likely to already have an existing schema), whereas in Experiment 6 participants are taught novel stems, in pairing with novel suffixes. Moreover, it is of interest whether sleep specifically provides an additional benefit to the generalisation of items derived from a set with greater variability.

Importance of stem variability for frequency

It is also of interest how increased stem variability interacts with the factor of suffix frequency. In Experiment 6 each suffix will be seen with a greater number of stems than in Experiment 5. In Experiment 6 there will be 12 different stems per suffix, as opposed to 8 as in Experiment 5.

There is some evidence suggesting that this greater variability makes the mapping easier to detect. For instance, Ford, Davis and Marslen–Wilson (2010) investigated form-meaning mapping regularities, whereby the regularity was contained within affix morphemes of English words (e.g. the regular affix "ly" can be applied at the ending of several words, such as "densely", "abruptly", "bluntly", "legally"). Lexical decision tasks were used to investigate how the frequency of this affix regularity is related to family size of words – i.e. the number of words which have the same affix in common. The affixes within the items used could be distinguished on their productivity. The more productive affixes (e.g. "ness") are frequently used in English to combine new words, whereas the less productive affixes (e.g. "the") are now used less frequently. The issue of productivity in affixes can also be related to the issue of consistency. The less productive affixes were less consistent in predicting the meaning of a word as they were combined with a greater number of words. For example, "age" could be applied at the ending of both verbs (e.g. "breakage"), and nouns (e.g. "orphanage"). Meanwhile, the more consistent and therefore more predictable affix "ness" could only be applied at the end of adjectives (e.g. "neatness").

In their study it was found that participants had faster reaction times for words with more productive affixes, than for words with less productive affixes. Base frequency – i.e. the frequency with which each stem appeared in the input- was found to be a good predictor of reaction times for words with more productive affixes, but not for words with less productive affixes. Family size – i.e. the number of words with the same morpheme in common – was found to predict the reaction times for both the more and less productive affixes. That is, participants responded faster to the words combined from affixes which are generally used more frequently and across a larger set of words. This suggests that affixes which appear more frequently in the input and are typically seen paired with a greater number of stems, may be easier to recognise, than the affixes which are seen paired with a smaller number of stems.

Moreover, Monaghan, Chater and Christiansen (2005), on the basis of corpus analyses, have suggested that phonological regularities may cue grammatical categories more reliably when contained within words which occur less frequently, than in words which occur more frequently. In Experiment 6 when stem variability is increased (but the overall number of items is kept constant), each individual word will appear in the input less frequently, in comparison to Experiment 5. Accordingly, this may make the suffix cue more reliable in indicating gender, and hence increase the likelihood of suffix-gender regularity extraction.

Aims and predictions

The aim of Experiment 6 was to investigate whether increasing stem variability will benefit the extraction of the suffix-gender regularity. As with Experiment 5, in Experiment 6 there will be no determiners in the input, and only 1 suffix per gender. It is expected that suffix-gender regularity extraction will be boosted as a result of greater stem variability. This prediction is supported by findings from the studies discussed above (Ford et al., 2010; Eidsvag et al., 2015; Gomez, 2002; Monaghan et al.,2005; Onnis et al., 2003;2004; Tamminen et al.,2015) which suggest that increased exemplar variability increases the salience of the regularity and hence makes it easier to acquire.

5.4.1 Method

The participant criteria, design, materials and procedure were the same as in Experiment 5. The key differences are outlined below.

Participants

As already mentioned, the participant criteria were the same as in Experiments 1 - 5. Fifty-two monolingual participants from the University of York participated in this study. Out of these, two participants did not return for the second session, and one participant did not complete one of the tasks and therefore were excluded from analysis. There were 24 participants in the wake group, with a mean age of 19.46 ± 0.98 [SD], out of which 5 were male and 2 were left-handed. There were 25 participants in the sleep group, with a mean age of 19.96 ± 1.72 [SD], 10 were male and 3 were left-handed.

Materials

As with Experiment 5, there were no determiners and a total of 2 suffixes (1 per gender). Stem variability was increased from 8 stems per suffix-gender mapping in Experiment 5, to 12 stems per suffix-gender mapping in Experiment 6. Increasing stem variability in Experiment 6 affected the training task. In both Experiments 5 and 6, there were a total of 384 items overall in the training task (as with Experiment 4). In Experiment 5, the 16 different words were repeated 24 times. In Experiment 6, the 24 different words were repeated 16 times. Half of the words were masculine in their mapping to the natural gender of the picture (and the paired suffix), and the other half were feminine. This means that there was a variability of 8 stems per suffix-gender mapping in Experiment 5, and 12 stems per suffix-gender mapping in Experiment 6.

Procedure

The only difference in procedure from Experiment 5, was that there was 1 generalisation task (as opposed to two as in Experiment 5), with a total of 16 items (half consistent, half inconsistent).

5.4.2 Results

Language tasks

Training task (Session 1)

As with Experiments 1-5, the word-picture matching task was used to train participants on the new language mappings. Both accuracy at the end of training, and correct RTs on target and noise trials separately were analysed, as in the previous Experiments, to provide an indication of the learning of the language. Also as with Experiments 1-5, two versions of the language were used to counterbalance for suffix-gender mappings (e.g. "*ool*" – "male", "*eem*" – "female" in one version, vice versa in the second version). It was expected that participants from both groups would reach similar levels of performance at the end of training, reflecting similar levels of familiarity with the language.

Accuracy analysis

To analyse accuracy at the end of training, a two way between subjects ANOVA was run with the independent variables of group (sleep or wake) and language (version one or two). The dependent variable was accuracy (proportion of correct responses) at the final block (block 8) of the training task.

There was no significant main effect of group ($F(1, 46) = .04$, $p = .834$), no significant main effect of language version ($F(1, 46) = .00$, $p = .987$) and no significant interaction between group and language ($F(1, 46) = .31$, $p = .584$). As shown in Figure 65, all participants reached a similar level at the end of training, as expected.

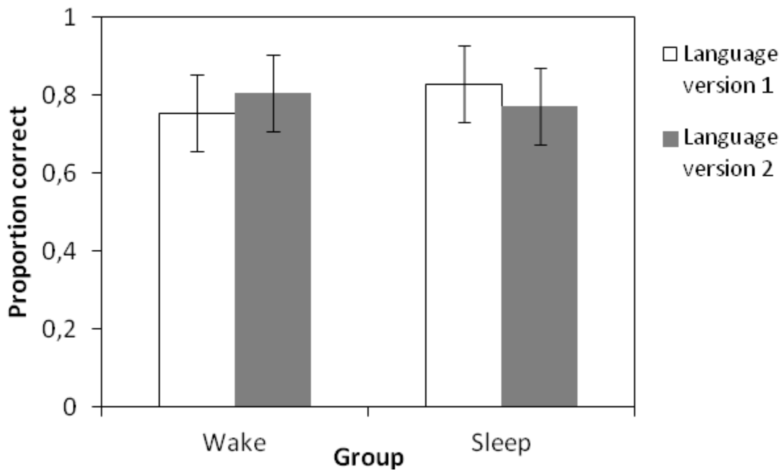


Figure 65. Performance at the end of training in Experiment 6.

Correct RTs analysis on the training task

The RT data for correct responses was analysed separately for target trials and separately for noise trials, as with previous Experiments. Across the multiple exposures to the word-picture mappings on the training task, participants should acquire the target mappings. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, 8 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake and sleep). First, the dependent variable was the change in target RTs compared to block 1.

There was a significant main effect of block ($F(1, 46) = 16.56$, $p < .001$), no significant main effect of group ($F(1, 46) = .05$, $p = .831$), and no significant interaction between block and group ($F(1, 46) = 1.06$, $p = .389$). As shown in Figure 66, participants from both groups had similar levels of improvement across the blocks, indicative of similar levels of learning of the language.

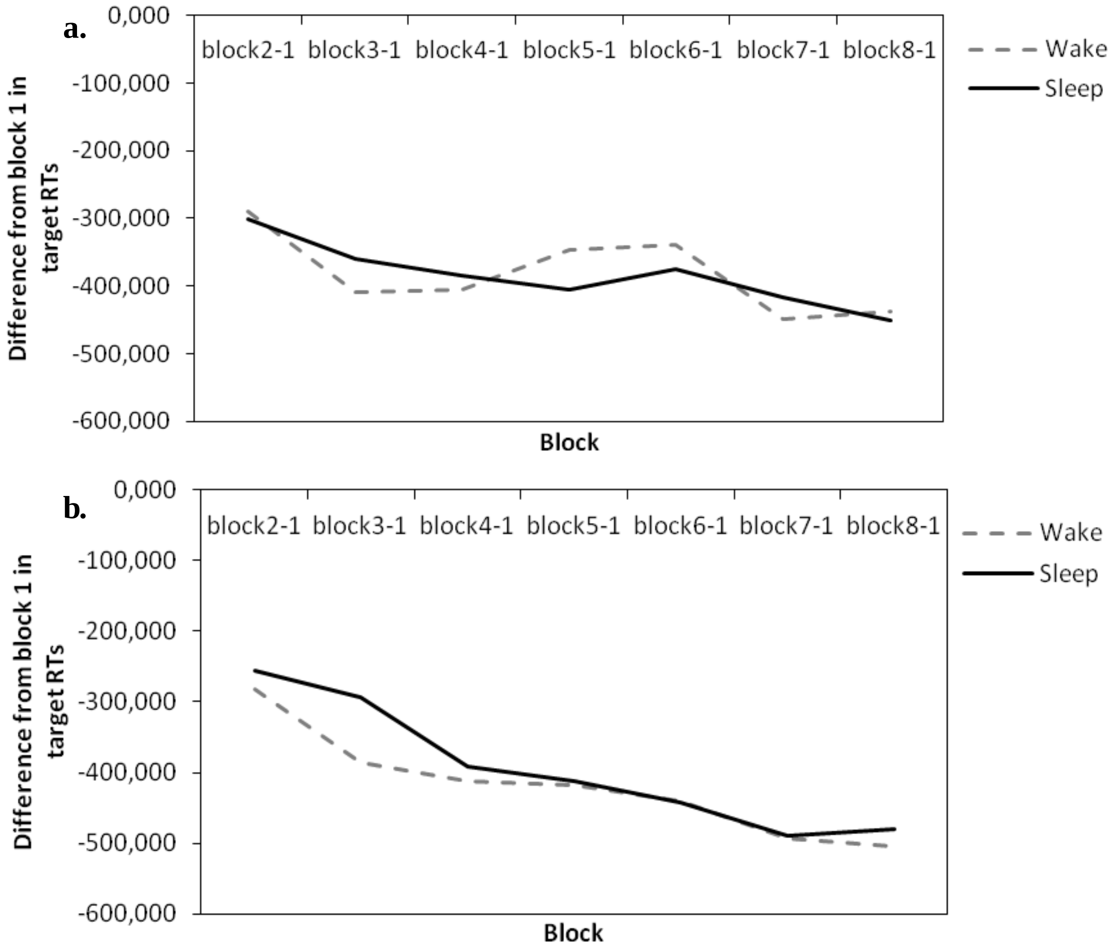


Figure 66. Difference from block 1 in target RTs in Experiment 6 (Figure 66a. shows language version 1 RTs, Figure 66b. shows language version 2 RTs).

There was also no significant main effect of language version ($F(1, 46) = .21$, $p = .651$) and no significant interaction between group and language ($F(1, 46) = .07$, $p = .794$). There was a significant interaction between block and language ($F(1, 46) = 2.31$, $p = .034$), and no significant interaction between block, group and language ($F(1, 46) = .24$, $p = .961$). Participants in language version 2 (Figure 66b) had greater improvement across the blocks than participants in language version 1 (Figure 66a).

Next, the difference from block 1 in noise RTs was analysed. One participant was removed from the analysis due to not having any correct noise RTs in all of the blocks. A mixed ANOVA was run, with the within subjects variable of block (difference between blocks 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, 8 and 1), the between subjects variable of language (version 1 or version 2) and the between subjects variable of group (wake and sleep). The dependent variable was the change in noise RTs compared to block 1.

There was a significant main effect of block ($F(1, 45) = 15.23$, $p < .001$), a marginally significant main effect of group ($F(1, 45) = 3.87$, $p = .055$), and no significant interaction between block and group ($F(1, 45) = .56$, $p = .764$). As shown in Figure 67, all participants improved across the blocks on noise RTs interaction (although the marginal effect of group indicates that participants in the wake group overall had a marginally greater RT difference between blocks 2-1 which was maintained throughout the training, compared to the sleep group).

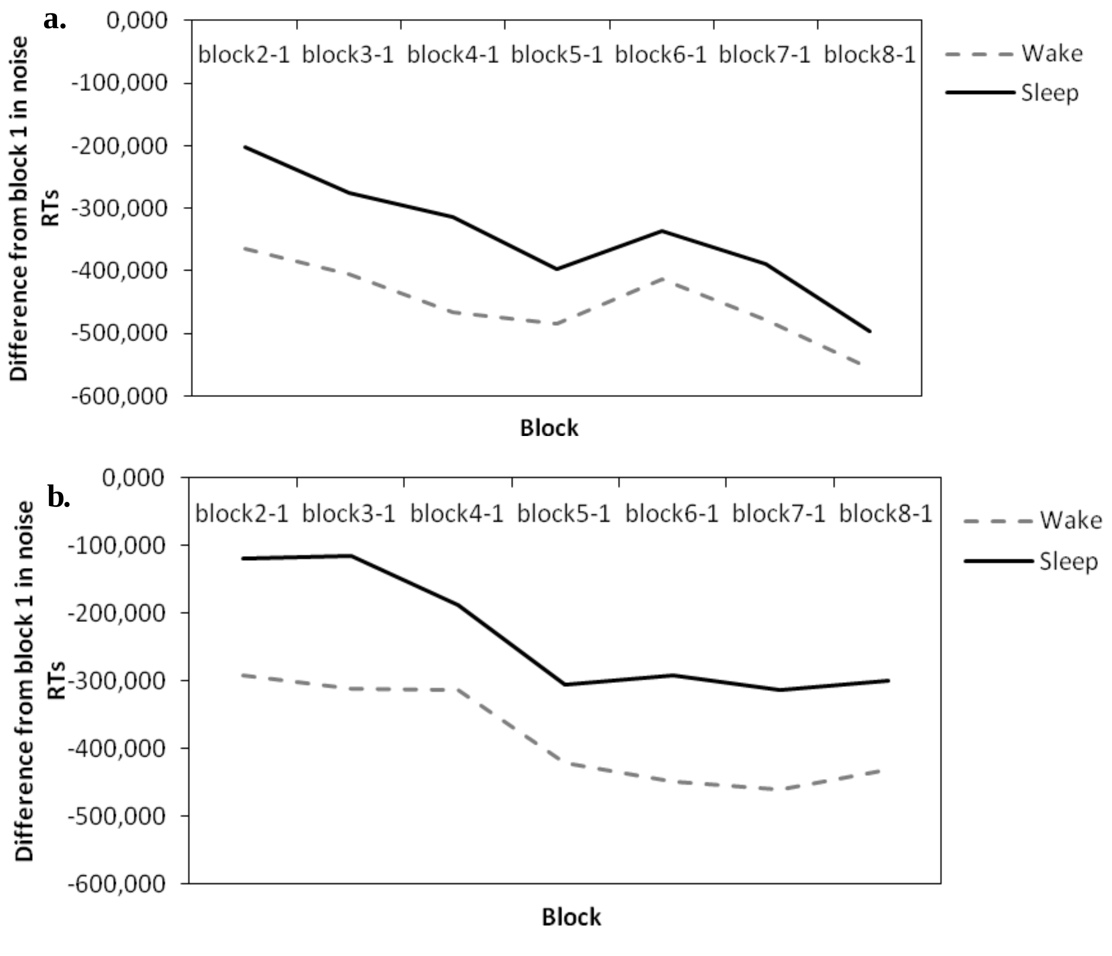


Figure 67. Difference from block 1 in noise RTs in Experiment 6 (Figure 67a. shows language version 1 RTs, Figure 67b. shows language version 2 RTs).

There was also no significant main effect of language version ($F(1, 45) = 1.92$, $p = .172$) and no significant interaction between group and language ($F(1, 45) = .10$, $p = .750$). There was a marginally significant interaction between block and language ($F(1, 45) =$

2.10, $p = .053$), and no significant interaction between block, group and language ($F(1, 45) = .27$, $p = .952$). Participants in language version 1 (Figure 67a) had marginally greater improvement across the blocks than participants in language version 2 (Figure 67b).

Overall, the findings from the accuracy analysis and the analysis of the correct target RTs training task suggested that both groups had similar levels of initial learning. Hence, any differences on the language tasks at session two are unlikely to be due to differences in initial acquisition, but rather due to differences in consolidation processes. This indication of differences between language versions will be further addressed in the General Discussion (i.e. Chapter 7), however given that the main focus of the investigation was on group differences, session two analyses will be carried out without taking language version into account.

Language tests (Session 2)

Most arbitrary mapping: vocabulary

Translation recognition task

As with Experiments 1-5, participants' memory of trained vocabulary was tested using a translation recognition task. Half of the items on this task consisted of correct, i.e. match individual word-picture mappings, the other half consisted of incorrect, i.e. mismatch individual word-picture mappings. Participants in the wake group had an accuracy rate of 75.3% for correct trials and 73.6% for incorrect trials. Participants in the sleep group had an accuracy rate of 81.0% for correct trials and 74.0% for incorrect trials. Hence, participants had an overall good level of recognition memory in terms of the accuracy.

A between subjects t-test was run with the independent variable of group (sleep or wake). A' discrimination was used as the dependent variable, in order to account for potential biases within participants' responses to correct and incorrect trials. There were no significant differences between the two groups ($t(47) = -.60$, $p = .551$).

As shown in Figure 68, all participants had similarly good discrimination between match and mismatch trials. As with Experiment 5 (when stem variability was lower), participants from both groups reached a similarly good level of performance on the recognition task.

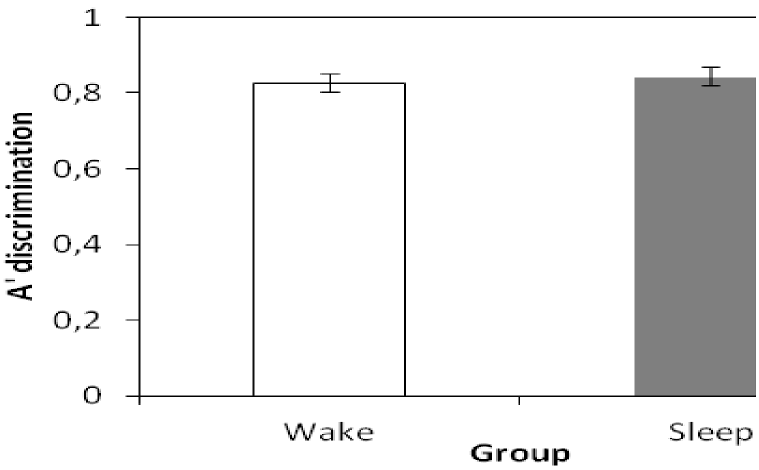


Figure 68. Performance on the translation recognition task in Experiment 6.

Recall (stems)

As with Experiments 1-5, the recall task was used to assess participants' memory for the trained word-picture mappings. Also in line with Experiments 1-5, recall of stems (as the most arbitrary mapping) was analysed separately from the recall of suffixes. Given that stems are the arbitrary mapping, it was expected that participants from the sleep group would have higher recall of stems than participants from the wake group, reflecting that sleep benefits the consolidation of arbitrary mappings.

It may be informative to provide a reminder of the stem recall rates in Experiment 5 (when stem variability was lower). For the sleep group, the average proportion of stems recalled correctly in Experiment 5 was .365 when both the picture and grapheme were present, and .352 when the picture only was present. For the sleep group in Experiment 6 (when stem variability was increased) it was .397 when both the picture and grapheme were present, and .282 when the picture only was present. For the wake group, the average proportion of stems recalled correctly in Experiment 5 was .256 when both the picture and grapheme were present, and .192 when the picture only was present. For the wake group in Experiment 6 (i.e. when stem variability was increased) it was .257 when both the picture and grapheme were present, and .163 when the picture only was present. These averages suggest that overall stem recall rates in Experiment 6 may have been quite similar to stem recall rates in Experiment 5 (when stem variability was lower).

To analyse stem recall accuracy, a mixed ANOVA was run with a between subjects factor of group (wake/sleep) and within subjects factor of cue type (picture and grapheme/ picture only). Stem accuracy (proportion correct out of total productions) was the dependent variable. There was a significant main effect of group ($F(1, 47) = 7.82$, $p = .007$) and a significant main effect of cue type ($F(1, 47) = 91.97$, $p < .001$). There was no significant interaction between group and cue type ($F(1, 47) = .95$, $p = .334$).

As Figure 69 shows, participants from both groups recalled a higher proportion of stems correctly when the grapheme cue was also present than when the picture only cue was present. More importantly, participants from the sleep group recalled a higher proportion of stems correctly than participants from the wake group, now that stem variability was increased.

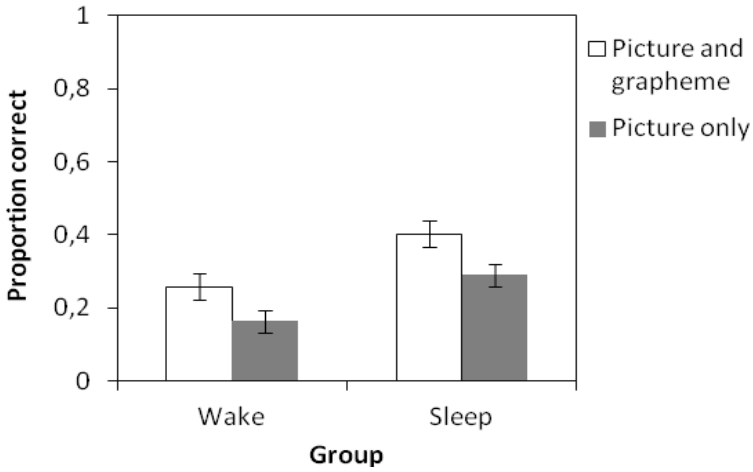


Figure 69. Stem accuracy on the recall task in Experiment 6.

To summarize, now that stem variability was increased, the results from the vocabulary tests provided evidence for a benefit of sleep for consolidation of individual mappings at the arbitrary level (i.e. stems) on the recall task, in line with Experiments 1 and 2. As with Experiments 1-5 there was no evidence for a benefit of sleep on the translation recognition task.

Systematic mapping: (suffixes)

Recall (suffixes)

As with Experiments 1-5, performance on the recall task was also used to compare the two groups on their suffix recall. It was expected that participants from the sleep group would have higher recall of suffixes at the level of the individual suffix-picture mapping, than participants from the wake group, in line with findings from the previous Experiments in this thesis.

It may be informative to provide a reminder of the suffix recall rates in Experiment 5 (when stem variability was lower). For the sleep group, the average proportion of suffixes recalled correctly in Experiment 5 was .735 when both the picture and grapheme were present, and .718 when the picture only was present. For the sleep group in Experiment 6 (when stem variability was increased) it was .670 when both the picture and grapheme were present, and .567 when the picture only was present. For the wake group, the average proportion of suffixes recalled correctly in Experiment 5 was .602 when both the picture and grapheme were present, and .570 when the picture only was present. For the wake group in Experiment 6 (i.e. when stem variability was increased) it was .470 when both the picture and grapheme were present, and .373 when the picture only was present. These averages suggest that overall suffix recall rates in Experiment 6 (when stem variability was increased) may have been slighter lower compared to suffix recall rates in Experiment 5 (when stem variability was lower).

To analyse suffix recall, a mixed ANOVA was run with a between subjects factor of group (wake/sleep) and within subjects factor of cue type (picture and grapheme/ picture only). Suffix accuracy (proportion correct out of total productions) was the dependent variable. There was a significant main effect of group ($F(1, 47) = 6.17$, $p = .017$), a significant main effect of cue type ($F(1, 47) = 39.03$, $p < .001$) and no significant interaction between group and cue type ($F(1, 47) = .04$, $p = .850$).

As Figure 70 shows, participants from both groups recalled a higher proportion of suffixes correctly when the grapheme cue was also present than when the picture only cue was present. More importantly, participants from the sleep group recalled a higher proportion of suffixes correctly than participants from the wake group, now that stem variability was increased.

Better suffix recall of the sleep group in this task might have been driven by either better individual suffix-picture mappings' memory, or by suffix-gender regularity extraction. The suffix generalisation task presented next will provide a further test of suffix-gender regularity extraction, and hence clarify whether the benefit of sleep on the recall task reflects individual suffix-picture mappings' memory of suffix-gender regularity extraction.

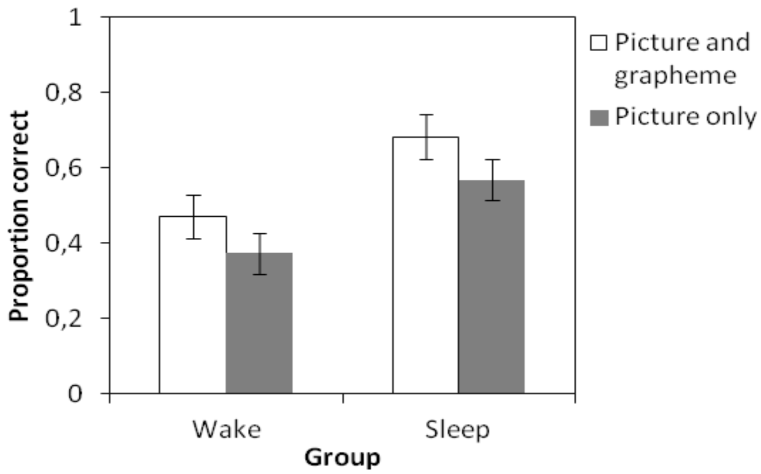


Figure 70. Suffix accuracy on the recall task in Experiment 6.

Suffix generalisation task

As with Experiments 1-5, to test participants' ability to apply the suffix-gender regularities they learnt during training to unseen items we used a suffix generalisation task. Experiments 1-5 have not provided strong evidence of the suffix-gender regularity extraction. In particular, in Experiment 5 (as with Experiment 6), there were also no determiners in the input, and suffixes were also fully systematic. Yet, there was still no evidence of suffix-gender regularity extraction in the sleep group, albeit some evidence of suffix-gender regularity extraction in the wake group. It was of interest whether now that stem variability was increased, there would be evidence of suffix-gender regularity extraction in both of the groups. As a reminder, in this task we are interested in the pattern of endorsement of consistent versus inconsistent items, as indication of generalisation ability.

A mixed ANOVA was run with the between subjects variable of group (sleep or wake) and within subjects variable of consistency (items consistent or inconsistent with trained regularities). Proportion endorsed (pressing 'match' to consistent and inconsistent items) was the dependent variable. There was a significant main effect of consistency ($F(1, 47) = 23.15$, $p < .001$), no significant main effect of group ($F(1, 47) = .91$, $p = .346$), and no significant interaction between group and consistency ($F(1, 47) = .22$, $p = .639$). This suggests that all participants extracted the suffix-gender regularity, regardless of group, as shown in Figure 71.

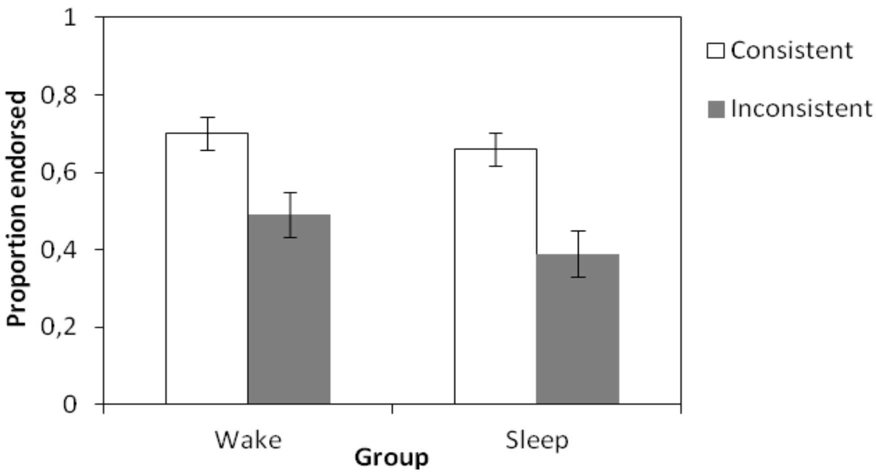


Figure 71. Suffix generalisation performance in Experiment 6.

One sample t-tests against chance (0.5) revealed that participants in the sleep group only endorsed consistent items above chance, but not inconsistent (consistent: $t(24) = 4.22$, $p < .001$, inconsistent: $t(24) = -1.60$, $p = .124$). Participants in the wake group also only endorsed consistent items significantly above chance, but not inconsistent items (consistent: $t(23) = 4.64$, $p < .001$), inconsistent: $t(23) = -.16$, $p = .874$). Hence, the tests against chance further confirmed that the pattern of endorsement was in line with suffix-gender regularity extraction.

As shown in Figure 71, all participants endorsed consistent items significantly more than inconsistent items, regardless of group. This provides evidence for suffix-gender regularity extraction, and no evidence for an additional benefit of sleep. This suggests that in the suffix recall task, the greater benefit of sleep (compared to wake) was based on the knowledge of the *individual* suffix-picture mappings at the arbitrary level, as opposed to the knowledge of the suffix-gender regularity.

To summarize, the suffix tests have provided evidence for a benefit of sleep for the consolidation of suffixes at the level of item learning (recall task) as with Experiments 1-3. There was no evidence that sleep benefited the extraction of the suffix-gender regularity (suffix generalisation task), as in the previous Experiments. Regardless of sleep, there *was* evidence for the extraction of the suffix-gender regularity, now that the stem variability was increased.

5.5 Discussion

The aim of Experiment 6 was to investigate whether increasing stem variability would boost the extraction of the suffix-gender regularity. Indeed, now that stem variability was increased, there was evidence that the suffix-gender regularity was extracted, regardless of sleep. The findings are in line with previous evidence suggesting that increased stem variability boosts the extraction of regularities (e.g. Gomez, 2002), which will be discussed further in the next section (i.e. Chapter 5 General Discussion).

Experiment 6 also provided evidence that sleep is beneficial for the consolidation of arbitrary mappings at the individual level (as evidenced by stem recall and suffix recall). Given that there was no evidence for a benefit of sleep for the extraction of the suffix-gender regularity (as measured by the suffix generalisation task), this suggests that the benefit of sleep on the suffix recall task was specific to consolidation of *individual* suffix-picture mappings. The lack of evidence for sleep benefitting extraction of the suffix-gender regularity, despite increased stem variability (as well as despite suffixes being fully systematic and the only cue to gender) is consistent with our predictions on the basis of the CLS model (e.g. McClelland et al., 1995).

5.6 Chapter Summary and General Discussion

The aim of Experiments 5 and 6 was to investigate whether the presence of determiners in the input also cueing the gender of the pictured characters, might have inhibited the extraction of the suffix-gender regularity in Experiments 1-4. It was also of interest whether suffix-gender regularity extraction would be boosted with increased stem variability. It was found that removing determiners from the input was not sufficient to result in suffix-gender regularity extraction. In other words, despite suffixes being a fully systematic mapping, removing the potentially competing cue (i.e. the determiners) from the input was not sufficient to increase the strength of suffixes as a cue to gender (to use the terms of the Competition model (e.g. Bates & Macwhinney, 1987)). Increased stem variability was additionally needed to boost suffix-gender regularity extraction, regardless of sleep.

Redundancy of multiple cues and suffix-gender regularity extraction

The findings from recall tasks in Experiments 1-4 have provided evidence that suffixes *are* noticed, as the individual suffix-picture mappings were learned (although not to a high level of accuracy). Despite some degree of individual suffix-picture mapping knowledge, Experiments 1-4 have not shown strong evidence of suffix-gender regularity extraction. The extraction of suffix-gender regularities in Experiment 6 suggests that despite suffixes being at the end of the word (hence making them less salient, in comparison to determiners, as we have previously suggested), suffixes *do* have the potential to not only be noticed (i.e. as individual suffix-picture mappings), but to also be processed for regularity extraction. Our findings suggest that in order for suffixes to be processed for extraction, removing competing cues (i.e. the determiners) from the input is *not* sufficient, despite suffixes being a fully systematic mapping. Increased stem variability is also needed for extraction.

As already discussed, there is evidence that when there are multiple cues to the same meaning, the two cues compete with each other (e.g. Bates & MacWhinney, 1987; MacWhinney, 1987), making one of the cues more difficult to acquire, due to its redundancy (e.g. Robinson, 2002). Overall, it was expected that suffix-gender regularities would be extracted without the presence of determiners in the input. We suggested that determiners might have inhibited suffix-gender regularity extraction, due to determiners being potentially perceived as the more salient and stronger cue, guiding attention away from suffixes. It was expected that by removing determiners from the input, the learner's attention would be guided to suffixes. Yet, purely removing determiners was not sufficient for suffix-gender regularities to be extracted.

In order to clarify these results, two further Experiments would be needed. In one of these Experiments, stem variability would be increased in the *presence* of determiners (while in Experiment 6 we increased stem variability with no determiners in the input). In the second experiment, stem variability would be increased (without determiners in the input), and the suffix-gender regularity would be made less systematic in comparison to Experiment 6 where suffixes were a fully systematic mapping. This would be done by increasing the number of suffixes from 1 per gender (as with Experiment 6), to 2 per gender. The results from such Experiments would shed light on the extent to which the presence of determiners inhibits suffix-gender regularity extraction, and the extent to which systematicity is important for suffixes, or whether stem variability is the *only* crucial factor for suffix-gender regularity extraction.

In addition, in the context of the already discussed IP model (e.g. VanPatten, 2004), the suffix cue was redundant in our Experiments 1-4 not only because gender could already be derived from the determiner cue, but also because it is of lower priority for acquisition of meaning than stems. This suggests that the priority of acquisition of the grammatical form could be increased by increasing the number of categories within the input, beyond just the one category of gender.

Finally, having considered the disadvantages of redundancy, it is important to consider its potential advantages for language learning. This potential advantage of learning is considered in the context of Bahrick, Lickliter and Flom's (2004) interesensory redundancy hypothesis. Bahrick et al. (2004) outline that *intersensory redundancy* arises within a multimodal input, which requires more than two senses to be processed, with a degree of overlap. Conversely, intersensory redundancy does *not* arise in a unimodal input, due to its nature. Crucially, this intersensory redundancy (in a multimodal input) can be advantageous for learning, in terms of the way it influences the learner's attention. It is crucial that the learner's attention is directed to the most meaningful aspects of the input, in order for the information acquired to be relevant for the learner. The presence of interesensory redundancy highlights the salience of the amodal aspects of the input. Therefore, the learner's attention is directed to the amodal aspects first (prior to other aspects of the input). Therefore, the learning of amodal aspects is more robust in the presence of intersensory redundancy.

Similarly, Monaghan (2017) found that in language learning (specifically, in the context of word-natural referent mappings), the presence of multiple cues to the *same* function can be beneficial in the learner's task of understanding what a specific, individual word means (i.e. in figuring out the correct word-natural referent mappings). This is because not of all the cues (e.g. cues to category or function in our Experiments) are always available in the input. Consequently, if the learner is exposed to multiple cues to function instead of just one cue to function, the learner does not have to entirely rely on just *one* cue to determine the word meaning. Therefore, even if a particular cue is not available in the input, the learner will still be able to understand the meaning of the word, by using another cue (to the same function). This makes learning more robust, particularly when considering the variety in language, i.e. the likelihood that a particular cue to function might not always be available. On the other hand, as found by Monaghan (2017), having more cues to the same function within the input, means that the learner has more information to process. This is disadvantageous, as it puts a "strain" on the learner's cognitive resources. Therefore, a "trade-off" arises, between the advantage of being more robust in the face of language variability, and the disadvantage of the "strain" placed on the learner's cognitive resources. Further research is required to determine at exactly which point this "trade-off" occurs (i.e. at which point the advantage becomes greater than the disadvantage).

Consequently, it is *possible* that the reason that we did not see evidence of suffix-gender regularity extraction despite eliminating the determiner-gender regularity, is because simultaneously as removing some of the disadvantage of this redundancy, we also removed some of its advantage. The learners *may* have found it more difficult to comprehend the Experimental words without this redundancy, as the redundancy may have made the more meaningful aspects more salient, as suggested by Bahrick et al. (2004). This suggestion requires further research. However, when considering the advantages of redundancy within our results, it is also crucial to remember that in our Experiments we investigated the *initial* stage of learning. In the long term, as learners gain more proficiency, they may indeed be benefited by having multiple cues to the same function (in this case both determiners and suffixes). However, during the initial stage of learning, when the mappings have not yet been acquired to a high level (as evidenced by our session two tasks), the cognitive strain in acquiring the mappings may still be so high, that at this *initial* stage of learning, having multiple cues may provide more of a disadvantage than an advantage to learning. It may be that, as found by Tamminen et al. (2015), at this *initial* stage of learning, one mapping needs to be learned alone, and allowed opportunity for consolidation, before introducing the second mapping, in order for both mappings to be acquired.

The effect of stem variability and explanations for suffix-gender regularity extraction

From our findings, it is still unclear whether the extraction of suffix-gender regularities is available immediately and purely dependent on the language input (such as stem variability), or whether it develops following a consolidation delay only, and is not available immediately. This idea that regularity extraction may depend on the language input and hence is available immediately, as opposed to following a consolidation delay, is also supported by the findings from Experiment 5. Specifically, the findings suggested that the extraction of the suffix-gender regularity may be dependent on the initial level of learning. As a reminder, in Experiment 5 there was evidence of suffix-gender regularity extraction in the wake group, with indication that this may have been driven by the wake group's better performance at training (as evidenced by the consistency x group x task order x session 1 training covariate interaction on the suffix generalisation task, as discussed in the Experiment 5 Discussion section).

There are also previous findings that increased variability boosts regularity extraction immediately, as previously found by Gomez (2002), Onnis et al., (2003; 2004) and Eidsvag et al., (2015). In addition, as found by Tamminen et al., (2015) regularity extraction is also still observed after a one-week delay. The above studies have provided evidence suggesting that it is increasing the middle element or root variability, as opposed to increasing repetition or overall frequency, which boosts detection of the regularity contained within the mapping. Our findings are in line with the above studies, suggesting that as a result of increased stem variability, suffix-gender regularity extraction is observed after a 12-hour consolidation delay.

There may be several angles from which to explain the reason why increased stem variability may have resulted in increased suffix-gender regularity extraction. Firstly, according to Gomez (2002)'s explanation, these findings may suggest that increasing stem variability increases the salience of the *systematicity* of the suffixes. This makes the suffix-gender regularity easier to detect. As already mentioned, our results are in line with Gomez (2002)'s findings that when variability is lower, participants focus on adjacent dependencies (i.e. the individual mappings), while when variability is higher, participants focus on the regularities contained within the non-adjacent dependencies.

Yet, increased salience of the *systematicity* of the suffixes may not be the only explanation for suffix-gender regularity extraction. To elaborate, despite there being a greater number of stems to consolidate in Experiment 6, than in Experiment 5, stem recall was similar in both Experiments. Meanwhile, suffix recall was lower in Experiment 6 (with increased stem variability), than in Experiment 5. This suggests that increasing stem variability guided the learner's attention away from *individual* suffixes, to stems (otherwise, stem recall would have been lower in Experiment 6 (increased stem variability) than Experiment 5). This increased attention to stems, resulted in lower consolidation of *individual* suffixes, and in turn boosted the extraction of the suffix-gender *regularity*. Hence, increasing stem variability may have also increased *individual* stem salience, not only the salience of the suffix *systematicity* (as suggested by Gomez, 2002).

It is interesting to consider further whether the extraction of the suffix *systematicity* is affected by *individual* suffix memory. Indeed, in Experiment 6, when suffixes were extracted, *individual* suffix memory was lower (than in Experiment 5). There are previous findings suggesting that individual item memory may block generalization of the overlapping regularities (e.g. Vlach et al., 2012; Werchan & Gomez

2014). It is possible that when stem variability was lower (in Experiment 5), the higher individual suffix memory blocked the extraction of the suffix-gender regularities. When increased stem variability resulted in lower individual suffix memory, extraction was increased. This suggests that some forgetting of individual items may be needed for extraction to take place.

Another explanation for suffix-gender regularity extraction is provided from findings by Ford et al. (2010), and Monaghan et al., (2005) about frequency. According to these findings, suffix-gender regularities were more likely to be extracted in Experiment 6 (when stem variability was higher) than in Experiment 5 (when stem variability was lower), because of the factor of frequency. Increasing stem variability increases the frequency of *use* of each suffix per stem, given that the same suffix-gender regularity is seen across an increased number of stems. It also decreases the frequency of each individual word in the input.

This is important in light of usage-based theories of language acquisition (and associative learning of language), given that within such theories learning is exemplar-driven (e.g. Ellis, 2006). The increased number of exemplars in the input which all share the same suffix-gender regularity benefits schema construction. This is because a greater number of exemplars are available from which to parse out the regularity. This strengthens the resulting memory representation of the form-meaning mapping (such as the suffix-gender "construction"). Consequently, it appears to the learner that the *systematicity* of the suffix cue is more reliable in predicting gender in language in *general*, as opposed to in just a small sample of exemplars, making regularity extraction more robust (Bybee & Thompson, 2000).

However, a potential consequence of this increased frequency of suffix-gender regularity use (i.e. the same regularity being applied to a greater number of exemplars) is "autonomy" (Bybee, 2007). Given that the form-meaning mapping is represented autonomously, parsing the individual suffix from the stem becomes more difficult. This is supported by our finding that increased stem variability may have resulted in decreased individual suffix recall.

To summarize, suffix-gender regularities may have been extracted in Experiment 6, because increased stem variability may have increased stem salience, increased salience of the systematicity of the suffixes, increased forgetting of the individual suffixes, and increased frequency of use for suffixes, strengthening regularity schema and representation.

The role of sleep

Consistent with findings from Experiments 1 and 2, and predictions based on the CLS (McClelland et al., 1995), there was evidence from Experiment 6 to suggest that sleep is beneficial for consolidation of the individual mappings at the arbitrary level (stems), when stem variability was increased. This was not the case in Experiment 5, when stem variability was lower.

There was no evidence that sleep was beneficial for the extraction of the suffix-gender regularities, regardless of stem variability. In the discussion section of Chapter 2, we compared the findings from our Experiment 1 to Lau et al.'s (2011) study. In our Experiment 1 we did not find any evidence that sleep benefits the extraction of regularities, which was not in line with Lau et al.'s (2011) study. We suggested that one of the key differences was in *variability*. Namely, the stimuli in Lau et al.'s (2011) study were Chinese characters which belonged to *seven* semantic groups, whereas the items in all six of our Experiments (described in this thesis) mapped onto only *two* semantic categories (male and female). We suggested that increased variability may facilitate sleep-dependent benefits for extraction of regularities, whereas decreased variability may boost sleep-dependent benefits for consolidation of individual specific items. While we did not increase variability of the semantic categories, Experiment 6 has provided no evidence that increased variability of exemplars in the input (i.e. increased stem variability) boosts sleep-dependent benefits for extraction of regularities. Further research is still needed to investigate whether this finding also extends to increased variability of semantic categories.

Overall, despite its benefit for the consolidation of arbitrary mappings, the role of sleep in learning and consolidating systematic aspects of language mappings seems to be limited. What appears to be more crucial than sleep (especially for the extraction of systematic regularities within language mappings) is the language input itself. More specifically, the factors of cue position (beginning or end of the word), cue redundancy (and competition), cue systematicity and cue frequency (influenced by stem variability) appear to be crucial for the extraction of language mapping regularities – not sleep.

In addition, aside from sleep, more time may be required for an individual cue to be consolidated, before the next individual cue can also be consolidated (e.g. as previously found by Tamminen et al., 2015). The role of sleep, while important, is neither sufficient in itself, nor is the most crucial factor for facilitating language acquisition and consolidation. Sleep is just one of *many* important factors that affect these processes. However, particularly given that the updating of an adults' lexicon and grammar is a continuous and ongoing process (e.g. Bybee, 1998b; Heine & Kuteva, 2007), there is *a* role for sleep to play in language learning. More specifically, sleep may particularly play a role in the learning of new phonological forms (e.g. Dumay & Gaskell, 2007), and the learning of new words, i.e. new arbitrary form-meaning mappings (e.g. Williams & Horst, 2014).

Conclusions

Regardless of sleep, the lack of potentially competing and inhibiting determiners in the input and full suffix systematicity was not sufficient to result in suffix-gender regularity extraction. Increased stem variability boosted suffix-gender regularity extraction, at the expense of lower individual suffix memory. Increasing stem variability potentially made suffix-gender regularity extraction easier due to increased salience of the systematic suffixes, or due to increased frequency of *use* of the suffixes (as the same suffix-gender regularity could be "used" across a greater number of exemplars). Alternatively, increased stem variability may have increased individual stem salience, thereby leading to lower individual suffix memory, suggesting that some individual item forgetting is necessary for boosting extraction.

Moreover, the findings from Experiment 6 have provided evidence to confirm that sleep is beneficial for the consolidation of arbitrary mappings, but no evidence that sleep is beneficial for extraction of regularities, consistent with Experiments 1and 2, and the predictions of the CLS model (McClelland et al., 1995). This suggests that language input factors such as systematicity, exemplar variability and cue redundancy may be more crucial for the learning of the systematic mappings in language than sleep.

Chapter 6: The relationship between the arbitrary and systematic mappings and the declarative/procedural model

6.1 Introduction

In addition to the novel language, we also included a procedural and a declarative memory task in our Experiments. This was of interest two reasons. Firstly, the role of the procedural and declarative systems in language acquisition was of interest. Secondly, the role of sleep in the consolidation of procedural and declarative systems was of interest.

Ullman's declarative – procedural model

As discussed in Chapter 1, Ullman's declarative-procedural (DP) model (e.g. Ullman, 2001) makes predictions about how aspects of language *might* be related to the domain-general declarative and procedural memory systems. The two memory systems underlie different types of information and are also subserved by different neural systems (i.e. brain areas). In general, the declarative memory system stores memories about facts (including semantic memory) and events (including episodic memory). Declarative memory is typically subserved by the hippocampal brain areas and the medial temporal lobes, with memories eventually being stored in the neocortical areas (Ullman, 2001, 2004). Procedural memory meanwhile subserves not only cognitive skills (such as aspects of grammar learning), but also motor memory, actions and skills, and is typically thought to be acquired implicitly (e.g. Eichenbaum & Cohen, 2001). The procedural memory system is typically subserved by basal ganglia and frontal lobe structures (e.g. Ullman, 2001). Learning within the declarative system occurs relatively more quickly, whilst learning in the procedural system requires a longer time period with multiple exposures (e.g. Knowlton & Moody, 2008).

Ullman's DP model (e.g. Ullman, 2001) predicts that in the learner's first language, the learning of new phonological forms and form-meaning mappings can be linked to the declarative domain-general memory system. While in the learner's *first* language the learning of aspects of grammar may be linked to the domain-general procedural system, this may not be the case in the learner' *second* language. Specifically, given that the procedural system requires additional repeated exposures and time, the DP model predicts that initial stages of second language learning will be related to the declarative system.

Indeed, as already discussed in Chapter 1, there is some evidence to support the predictions that the domain-general systems might be related to aspects of language learning and consolidation. Specifically, Ettlinger, Bradlow and Wong (2012) have shown that the learning of simple grammatical patterns can be linked to the procedural system, whereas the learning of more complex, analogy based patterns can be linked to the declarative system. Similarly, Brill-Schuetz and Morgan-Short (2014) have shown that the consolidation of grammatical regularities regarding the word order in an artificial language sentence can be linked to the procedural system.

However, this may depend on the time course of learning. For instance, Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter and Wong, (2014) using the same artificial language have shown that early in learning, the learning of grammatical rules is linked to the declarative system (measured by a novel word translation task and also by measuring nonverbal memory of abstract images), whereas with time it becomes linked to the procedural system (measured by both the weather prediction task and the tower of London task).

In terms of the underlying neural systems, Breitenstein et al. (2005) previously found the learning of novel, arbitrary word-picture mappings (e.g. "bini – book") to be related to the hippocampal system. As the declarative system is in part subserved by the hippocampus as well (e.g. Ullman, 2001), this suggested a link between the learning of arbitrary mappings and the declarative system. This suggestion was also supported by Lieberman, Chang, Chiao, Bookheimer and Knowlton's (2004) findings that hippocampal activity (and medial temporal lobe activity) was linked to the retrieval of trained item chunks regardless of grammaticality, whereas the caudate nucleus (in basal ganglia) was linked to the retrieval of implicitly learned grammatical rules.

These findings can be related to domain-general memory subsystems according to which the caudate nucleus subserves the procedural system, whereas (as already mentioned) the hippocampus and medial temporal lobe subserve the declarative system (Ullman, 2001). The findings suggest a link between the retrieval of exact item chunks and the declarative system, and the retrieval of grammatical regularities and the procedural system.

The above studies suggest that the learning, consolidation and retrieval of individual, arbitrary items (such as our stems) might be linked to the declarative system. Meanwhile, the learning of grammatical regularities (such as our determiner-gender mappings, and suffix-gender mappings) might be linked to the procedural system (as with Brill-Schuetz and Morgan-Short, 2014). Alternatively, the learning of grammatical regularities might initially be linked to the declarative system, but with increased proficiency become linked to the procedural system (e.g. Morgan-Short et al., 2014).

Robertson's (2009) model

The relationship between language learning and memory systems is further complicated by the fact that inteference may arise between the two memory systems, as suggested by Robertson's (2009) model of competition between memory systems. This model is in part based on Brown and Robertson's (2007) findings that sleep (compared to the equivalent time awake) may protect memories from interference. More specifically, during wakefulness the acquisition of declarative information inhibits the consolidation of prior procedural information and vice versa (i.e. the acquisition of procedural information inhibits the consolidation of prior declarative information). Sleep may protect the consolidation process from inhibition as a result of interference.

Robertson's (2009) model concerns itself with how memories are processed post-encoding. Neural circuits remain activated even post encoding, affecting the processes of consolidation. The exact nature of this varies depending on the brain state at the time of consolidation, i.e. wake or sleep. The sleep state results in the disengagement of the two memory systems, decreasing the interaction, and hence supposedly decreasing the amount of interference between the procedural and declarative system.

During the wake state, there is bidirectional communication between the hippocampus and the neocortex. Meanwhile, during the sleep state, this communication is reduced, such that the neocortex communicates with the hippocampus unidirectionally. Consequently, sleep may protect memories from interference, by disengaging the communication that results in interference. On the basis of this model, we can predict that sleep might have a protective effect against loss of information (in comparison to the equivalent time awake) in the face of interference (as both a procedural and a declarative task were used).

6.2 Exploratory Investigation of declarative and procedural memory

In order to explore the relationships between different memory systems and language learning in all six Experiments described in previous Chapters, we also included declarative and procedural memory tasks (as previously already mentioned). To test procedural memory, we used a finger-tapping task, which was taken from Walker, Brakefield, Morgan, Hobson and Stickgold (2002) and Walker, Brakefield, Seidman, Morgan, Hobson, and Stickgold, (2003). To test declarative memory, we used a 2D object location task taken from Rasch, Buchel, Gais and Born, (2007). Both of these tasks have previously shown sleep-related memory consolidation effects (e.g. Rasch et al., 2007; Walker et al., 2002; 2003). Compared to the equivalent time awake, improvement is often found following sleep in procedural memory tasks (for example as with Walker et al., 2002, 2003), but only decreased forgetting following sleep in declarative memory tasks (for example as with Lahl, Wispel, Willigens, & Pietrowsky, 2008; Plihal & Born, 1997; Rasch et al., 2007; Tucker et al., 2006).

Aims and predictions

The aim of this Chapter was to firstly investigate the role of sleep versus wake dependent consolidation of procedural and declarative information. Secondly, it was to investigate the extent to which the consolidation of the declarative task and procedural task information correlates with aspects of the language tasks. As we are not testing declarative or procedural information in isolation, interference may arise either at acquisition and consolidation, or at retrieval. As a result of this interference, we may not observe the sleep-dependent benefits on the declarative and procedural tasks that have been previously observed.

Additionally, the data from the procedural and declarative tasks was correlated with the language measures from our Experiments 1-6. This is an exploratory analysis, based on the declarative-procedural model (e.g. Ullman, 2005). The findings from our previous chapters have shown that the stem recall and the suffix recall represent arbitrary mappings. It is possible that we may observe positive correlations between these arbitrary measures (namely, stem recall and suffix recall with the declarative task).

6.2.1 Method

Participants

As has been previously mentioned, participants across all of our 6 Experiments also completed the procedural tasks and the declarative tasks. This meant that the correlations with language tasks were analysed on the basis of 268 participants for the procedural task and 270 participants for the declarative task. The 16 participants from

Experiment 2 whose overnight polysomnography data were collected were used to analyse correlations between the tasks and the sleep stages.

Due to participants not completing the task, data from 4 participants in the procedural task was missing (1 participant in Experiment 3, 1 participant in Experiment 4, 2 participants in Experiment 6). Also due to participants not completing the task, data from 3 participants in the declarative tasks was missing (2 participants from Experiment 4 and 1 participant from Experiment 6).

The procedural task analysis across Experiments was computed on the basis of 124 participants in the wake group and 126 participants in the sleep group.The correlations with sleep stages (using data from Experiment 2) was computed on the basis of 16 participants. The declarative task analysis across Experiments was computed on the basis of 127 participants in the wake group and 126 participants in the sleep group. The correlations between procedural memory and the language measures were computed on the basis of 268 participants. The correlations between declarative memory and the language measures were computed on the basis of 270 participants. The exception to this, were the correlations of the declarative and the procedural tasks with the determiner and suffix generalisation task. These correlations were computed on the basis of 172 participants (this is because this task was used to measure determiner-gender regularities, and determiners were removed from the input in Experiments 5 and 6).

Design

As the data from the procedural and declarative tasks was collected across the 6 Experiments described in Chapters 2-5, the design was between-subjects, i.e. participants were either in the wake group, or in the sleep group, as already mentioned in the previous Chapters. All participants completed both the declarative task, and the procedural task.

Materials

Procedural memory task (finger-tapping task, previously used by Walker et al., 2002, 2003)

Keyboard keys (on a QWERTY keyboard) Z-V were labelled 1-4 (in order from left to right). There were two possible 5-digit string sequences which could be presented on screen: 4-1-3-2-4 and 2-3-1-4-2. One of the sequences was used as the training sequence, and the other untrained sequence was used to test for non-sequence specific motor improvements in session two. For example, if 4-1-3-2-4 was used as the training sequence, then during Session 1 the sequence 4-1-3-2-4 was repeated for 12 trials (with the first trial considered a practice sequence) during the exposure phase. At a later stage during Session 1, the sequence 4-1-3-2-4 was repeated for another 4 trials as the training stage (used for analysis to compare performance between Session 1 and Session 2). During Session 2 the same sequence (4-1-3-2-4) was presented again for 4 trials as a retest and a different untrained sequence (2-3-1-4-2) was presented also for 4 trials. The use of the two finger-tapping sequences as trained or untrained (i.e. whether 4-1-3-2-4 was used as the trained sequence or as the second untrained sequence for each participant) was counterbalanced across participants.

During each trial, the tapping sequence was presented in the middle of the screen for 30 seconds – participants were asked to tap continuously for this time using the marked computer keyboard keys. Then there was a break for another 30 seconds. Participants then had to press the spacebar to carry on to the next sequence. Participants were asked to use all 4 fingers (excluding the thumb) of their non-dominant hand (i.e. right-handed participants used their left hand) to type the sequence on the keyboard as quickly and accurately as possible. This task assessed participants' learning and consolidation of procedural memories.

Declarative memory task (2D object location, Rasch et al., 2007)

The stimuli consisted of 15 picture pairs (of known animals or objects) presented on a grey grid. During exposure, all of the 15 pictures picture pairs were seen 3 times (across 3 blocks). Participants saw a grid of grey squares (5 x 6 squares) on screen. On each trial a picture appeared somewhere on the grid (randomly) in place of a grey square for 1 second. It was then followed by its matching picture (i.e. exactly the same picture but in a different place on the grid) while the first picture remained on screen for another 3 seconds. Then, the two matching pictures automatically were replaced with the grey squares again (such that participants saw a blank grid of grey squares) before the next trial began. The inter-trial interval was also 3 seconds. Participants were asked to memorise each matching picture pair's place on the grid one pair at a time.

During the training (Session 1) and delayed test (Session 2), participants saw a picture appear somewhere on the grid and were asked to use the mouse to click to the location (grey square) of its matching picture on the grid. Following each participant's response (regardless of accuracy), the correct picture location was revealed and remained on screen for 2 seconds. During the training, the same procedure was used for 3 blocks (15 pairs used once in each block). Performance on the last block was taken as a measure of the training performance and compared to the 1 block at delayed test. This task was used to assess participants' learning and consolidation of declarative memories.

Procedure

All participants completed both the declarative task first, and the procedural task second (followed by the language tasks third). This order remained constant both at session 1 training and at session 2 retest (as shown in Table 15).

Table 15. Procedure (Order of tasks).

Session1

Tasks	Stimuli
Vigilance	
1. Declarative	**2D Object location exposure and training**
2. Procedural	**Finger-tapping exposure**
3. Language Training	Repetition task
	Word-picture matching
4. Procedural	**Finger-tapping training**

Session 2

Tasks	Stimuli
Vigilance	
1. Declarative	**2D object location delayed test**
2. Procedural	**Finger-tapping retest**
	Finger-tapping generalisation*
3. Language Tests	Recall
	Determiner selection**
	Suffix generalisation, determiner and suffix generalisation***
	Translation recognition

*New sequence. **Not included in Experiments 5 and 6. ***Only one generalisation task included in Experiment 6.

Notes on data analysis

As already mentioned in Chapter 3, polysomnography data for one participant in Experiment 2 (overnight sleep study) was incomplete due to technical problems (i.e. electrodes detaching during the night). For this reason, the correlations with sleep parameters (with the declarative and procedural data) were conducted without this participant (i.e. on the basis of 16 participants). The behavioural correlations with language data (for the sleep group) across the Experiments were conducted with this participant included.

As already described in Chapter 5, there were 2 suffix generalisation tasks in Experiment 5. As both of the tasks were the same in their nature (only differing in the individual items), only the first of the two tasks was used for correlations from Experiment 5. This would be in line with our other Experiments having one suffix-generalisation task, particularly as the results of Experiment 5 (described in Chapter 5) indicated differences in performance across the two suffix generalisation tasks.
There was no determiner and suffix generalisation task included in Experiments 5 and 6; hence the correlation with this task was done on the basis of 3, not 5 Experiments.

6.2.2 Results

Procedural task

Between groups comparison across Experiments

Participants' procedural memory was tested using a finger-tapping task (taken from Walker, Brakefield, Morgan, Hobson & Stickgold, 2002). As in other studies using the same task (e.g. Walker et al., 2002, 2003), the number of correctly completed five-digit sequences (e.g. 4-1-3-2-4) for each 30 second trial was used as the dependent variable (thus taking into account both speed and accuracy). Average performance of the final four training trials (Session 1) was compared with average performance of the four retest trials on the same tapping sequence (Session 2).

In this task, as we were interested in post consolidation-delay performance, we did not include the two immediate groups from Experiment 4 in the analysis. As Experiment 2 did not have a wake group, we did not include the sleep data in this analysis either. Hence, the analysis was computed on the basis of 124 participants in the wake group and 126 participants in the sleep group.

A mixed ANOVA was run with a between subjects factor of group (sleep or wake), a between subjects variable of Experiment (Experiment 1, 3, 4, 5, 6) and a within subjects factor of test (training or delayed test). The dependent variable was speed. The statistical analysis yielded no significant main effect of group ($F(1, 240) = .78$, $p = .537$). There was a marginally significant main effect of test ($F(1, 240) = 3.06$, $p = .082$), and no significant interaction between test and group ($F(1, 240) = .01$, $p = .929$).

Additionally, there was no significant main effect of Experiment ($F(1, 240) = .34$, $p = .849$), and no significant interactions between group and Experiment ($F(1, 240) = .78$, $p = .537$) or between test and Experiment ($F(1, 240) = .91$, $p = .456$). There was also no significant three-way interaction between group, Experiment and test ($F(1, 240) = .59$, $p = .672$).

As shown in Figure 72, all participants across the 5 Experiments had similar performance (i.e. speed), regardless of group. There was a marginally significant effect of test, reflecting some forgetting between training and retest for all of the participants. That is to say that instead of a sleep-related gain in speed (as with Walker et al., 2002; 2003), or a protective effect of sleep to maintain procedural memory in the face of interference (Albouy et al., 2015), there was some marginal loss of procedural memory following both wake and sleep.

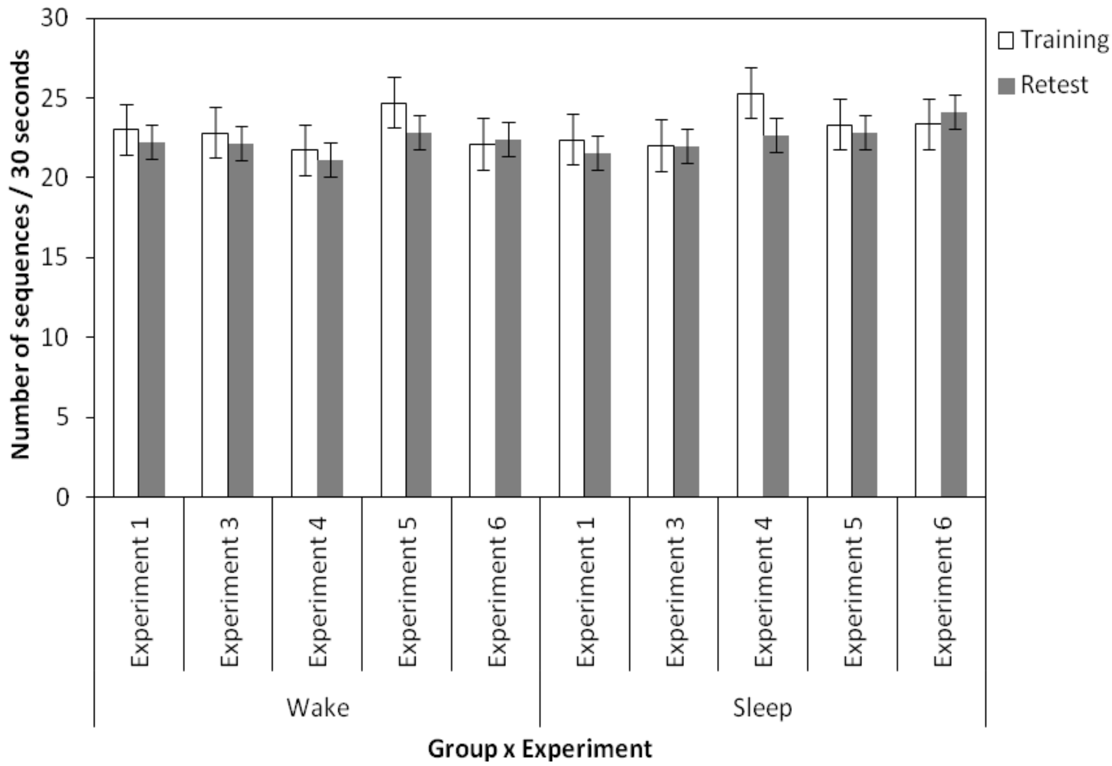


Figure 72. Performance on the procedural task across 5 Experiments.

In addition, we analysed procedural memory relative to session 1 performance. Namely, we subtracted performance at training from performance at test.A two-way between subjects ANOVA was run with a between subjects factor of group (sleep or wake), a between subjects variable of Experiment (Experiment 1, 3, 4, 5, 6). The dependent variable was the change in the number of correct sequences at test relative to training (i.e. performance at test – performance at training).

The statistical analysis yielded no significant main effect of group ($F(1, 241) = .01$, $p = .927$). There was no significant main effect of Experiment ($F(1, 241) = .92$, $p = .454$), and no significant interaction between group and Experiment ($F(1, 241) = .59$, $p = .669$). As with our previous analysis of the procedural data using a different dependent variable, there was no indication to suggest that sleep benefited procedural memory, unlike findings by Walker et al., (2002).

Given that we did not find an effect of sleep behaviourally (unlike Walker et al. 2002; 2003), we would not expect to find any correlations with sleep parameters, which was tested using the data from Experiment 2, presented below.

Performance across the two sessions on the procedural task

To run the correlation with sleep parameters, we used only the data collected in Experiment 2 (i.e the overnight sleep study), and therefore we firstly analysed the behavioural procedural data from Experiment 2 on its own. There was only one group of participants (the overnight sleep group with 17 participants) and no wake group tested in Experiment 2. Therefore, instead of an ANOVA, a within-subjects t-test was run on the variable of test (training or retest). As with Experiment 1, the dependent variable was speed (i.e. the average number of correctly completed sequences per 30 second trial).

The statistical analysis yielded no significant differences between the two tests ($t(1, 16) = .15$, $p = .879$). Hence behaviourally, as with our other 5 Experiments, there was no sleep related improvement in performance between training and retest. As shown in Figure 73, participants had similar performance during both training and retest.

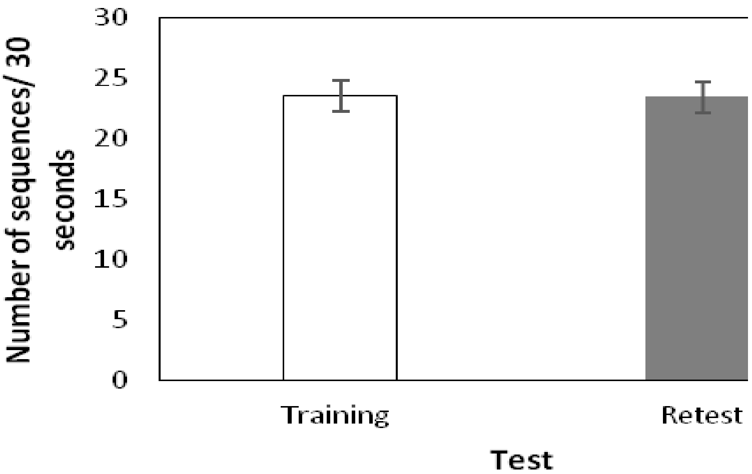


Figure 73. Experiment 2 Performance at training (Session 1) and retest (Session 2) on the finger tapping task.

Correlations with sleep parameters

One participant was taken out from the analysis due to missing sleep data, hence the analysis was based on 16 participants. Based on predictions from Fisher et al., (2002) and Wagner et al., (2003), we expected to find a significant correlation between the change in the correct number of sequences at test relative to training (i.e. procedural memory) and proportion of time spent in REM sleep. Alternatively, based on predictions from Walker et al., (2002), there could be a significant correlation with the proportion of time spent in stage 2 sleep. However, we also ran correlations with stage 1 sleep, SWS and sleep spindle count for thoroughness.

No significant correlation was found between procedural memory and proportion of time spent in REM sleep (r = -.083, p = .761). A marginally significant positive correlation was found between procedural memory and proportion of time spent in stage 2 sleep (r = .446, p = .084). Better preservation of procedural memory (at test relative to training) was associated with increased proportion of time spent in stage 2 sleep. There was also a significant negative correlation found between procedural memory and proportion of time spent in stage 1 sleep (r = -.598, p = .014). There was no significant correlation between procedural memory and proportion of time spent in SWS (r = -.066, p = .807), and also no significant correlation between procedural memory and sleep spindle count (r = .261, p = .347).

We have not found any evidence of consolidation related changes behaviourally, and no strong evidence of relationship with REM, inconsistently with our predictions. Stage 2 sleep (consistently with Walker et al., 2002) may play a role in procedural memory consolidation. In this case specifically, stage 2 sleep may play a role in the protective effect of sleep against procedural memory loss in the face of interference. Potentially, the negative correlation with stage 1 sleep could be explained when considering the *relative* proportions of time spent in each sleep stage. For instance, as the proportion of time spent in stage 1 sleep increases, the proportion of time spent in stage 2 sleep might decrease. This suggestion that stage 2 sleep might be important for procedural memory consolidation may need further exploration, as there was only marginally significant evidence for this correlation.

Declarative task

Between groups comparison across Experiments

Participants' declarative memory was measured using a visuospatial 2D object location task (Rasch, Buchel, Gais & Born, 2007). Performance at training (Session 1) was compared with the performance at the delayed test (Session 2). As with the procedural task analysis, the two immediate groups from Experiment 4 and the Experiment 2 participants (i.e. overnight sleep group only study) were not included in the analysis. Hence, (as already mentioned in the Participants section in the Method) the analysis was based on 127 participants in the wake group and 126 participants in the sleep group.

A mixed ANOVA was run with a between subjects factor of group (sleep or wake), a between subjects variable of Experiment (Experiment 1, 3, 4, 5, 6) and a within subjects factor of test (training and retest). The dependent variable was accuracy i.e. the proportion of correctly recalled card locations (based on the correct location clicks). If there is evidence of sleep-related consolidation, then the key interaction expected here would be between group and test. The direction of this interaction should reflect that there was less forgetting following sleep, than the equivalent time awake (e.g. Rasch et al., 2007).

The statistical analysis yielded a significant main effect of group ($F(1, 243) = 5.31$, $p = .022$), a significant main effect of test ($F(1, 243) = 50.59$, $p < .001$), and a significant interaction between test and group ($F(1, 243) = 16.68$, $p < .001$). As shown in Figure 74, while participants from both groups showed forgetting, this forgetting was greater in the wake group. Hence, sleep had a protective effect against loss of declarative information. (The performance of the sleep group was overall higher as well).

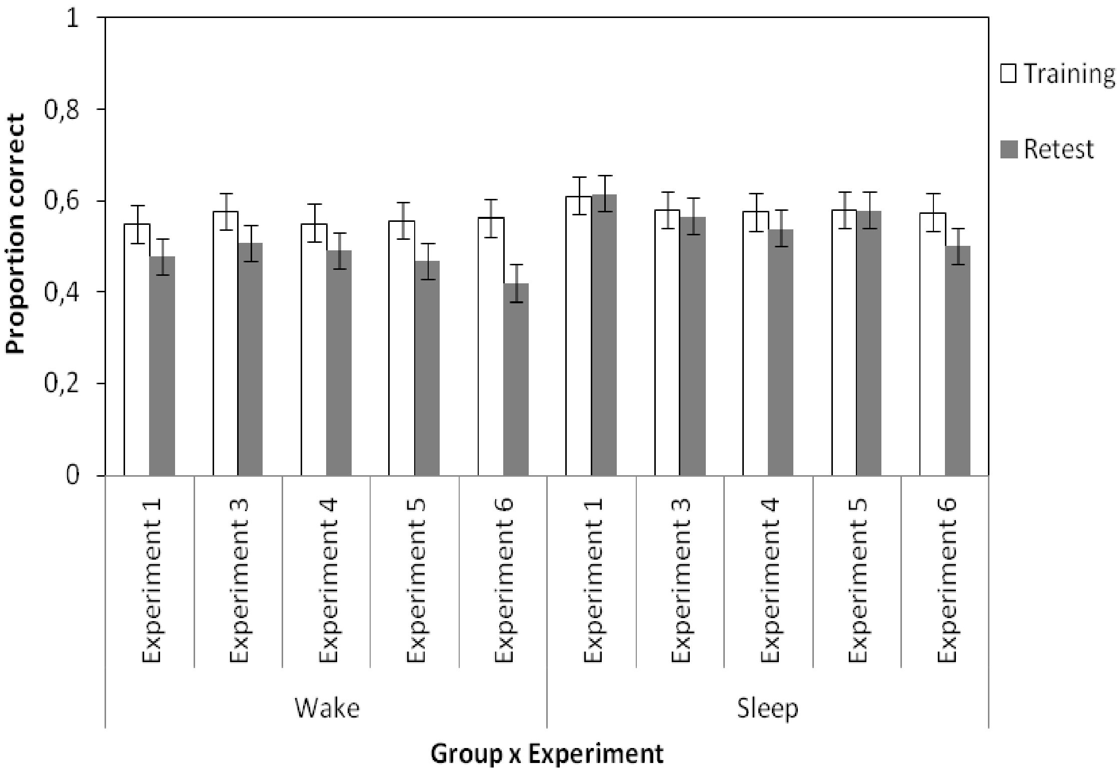


Figure 74. Declarative task performance across 5 Experiments.

Additionally, there was no significant main effect of Experiment ($F(1, 243) = .48$, $p = .750$), and no significant interaction between group and Experiment ($F(1, 243) = .27$, $p = .892$). There was a significant interaction between test and Experiment ($F(1, 243) =$

2.92, p = .022). There was no significant three-way interaction between group, Experiment and test (F(1, 243) = .54, p = .710).

As shown in Figure 74, regardless of group, forgetting between training and retest was the greatest in Experiment 6 and the least in Experiment 1, as evidenced by the significant Experiment x test interaction. In order to investigate whether the protective effect of sleep against further loss of declarative information could be related to SWS (as with Rasch et al., 2007), a correlational analysis was run.

Correlations with sleep parameters

In this declarative task analysis, we investigated whether there was a protective effect of sleep, driven by SWS, as found by Deliens, Leproult, Neu and Peigneux (2013) and for this particular task by Rasch et al., (2007). There was only one group of participants (the overnight sleep group) and no wake group tested in Experiment 2. Therefore, instead of an ANOVA, a within-subjects t-test was run on the variable of test (training and retest). As with Experiments 1 - 6, the dependent variable was accuracy. The statistical analysis yielded no significant differences between performance at training and retest (t(1, 16) = -.57, p = .579).

As shown in Figure 75, participants had similar performance both at learning (in session 1) and at retrieval (in session 2). In the analysis of Experiments 1-6 we found stronger forgetting in the wake group, than in the sleep group. Hence, if sleep has a protective function, this result in Experiment 2 would be expected, as there was no loss of memories between training and retest in Experiment 2. It could also be the case that individual differences in the amount of change from training to retest could correlate with the proportion of time spent in SWS. Yet, there was no significant correlation between proportion of time spent in SWS and the amount of change from training to retest (r = -.254, p = .342).

In addition, we also ran correlations between the amount of change from training to retest with other sleep parameters, for thoroughness. There was no significant correlation between proportion of time spent in stage 1 and the amount of change from training to retest (r = -.188, p = .501). There was no significant correlation between proportion of time spent in stage 2 and the amount of change from training to retest (r = -.172, p = .540). There was no significant correlation between proportion of time spent in REM and the amount of change from training to retest (r = .202, p = .470). Finally, there was no significant correlation between sleep spindle count and the amount of change from training to retest ($r = -.231$, $p = .408$).

Overall, as with the analysis of the declarative task between Experiments, in Experiment 2 sleep preserved declarative memory from being forgotten between training and retest, with no indication that this was mediated by SWS unlike our predictions, nor by any other sleep parameters.

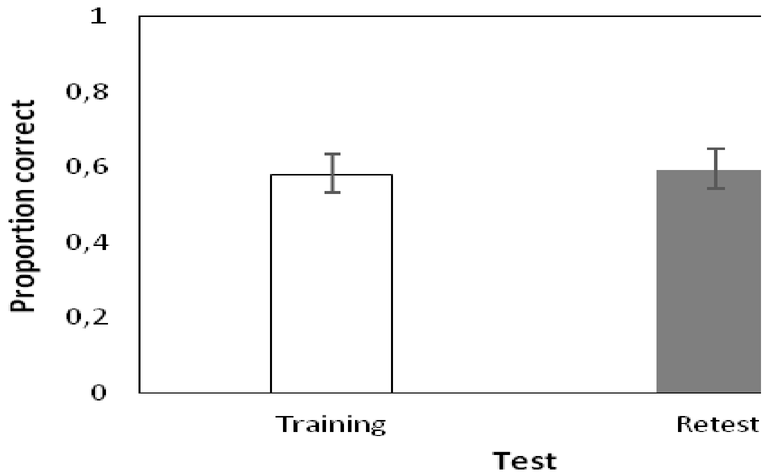


Figure 75. Performance on the declarative task in Experiment 2.

Correlations with language tasks

We ran correlations between the procedural and declarative tasks with the language tasks, in order to test the predictions of the declarative-procedural model (Ullman, 2005; Ullman, 2015) that arbitrary language mappings may be subserved by the declarative system. To elaborate, we are focusing on the test task correlations as the different language tasks at test could be viewed as assessing/ reflecting different aspects of language processing, according to the declarative-procedural model (e.g. Ullman, 2005). The Bonferroni correction was applied to account for multiple comparisons. Based on the 8 comparisons the *p*-value was set at .00625.

As shown in Table 16, there were significant (Bonferroni corrected) positive correlations between the session two declarative task performance and the two arbitrary measures, namely stem recall (both cue types) and suffix recall (both cue types). As a reminder, the suffix recall can be considered a measure of arbitrary mapping, because there was no evidence across experiments that suffix recall reflected the extraction of the regularity, but rather the memory of individual word-picture mappings. This suggests that performance on the declarative task was subserved by the same system as stem recall and suffix recall, consistent with predictions from the declarative-procedural model (Ullman, 2005; Ullman, 2015).

In addition, there was a significant positive correlation between session two declarative task performance and inconsistent suffix endorsement. The correlation with inconsistent suffix endorsement *may* reflect individual suffix knowledge, as opposed to suffix-gender regularity extraction. This is because participants who had extracted the suffix-gender regularity should have *low* levels of *inconsistent* suffix endorsement. Consequently, participants with *higher* levels of inconsistent suffix endorsement may be relying on individual phonological suffix knowledge, hence the positive correlation with the declarative task.

Table 16. Correlation coefficients for session 2 procedural retest, session 2 declarative retest with session 2 language measures.

Language measure	Procedural memory at test	Declarative memory at test
Stem recall (picture & grapheme)	.131*	.228***
Stem recall (picture only)	.125*	.249***
Suffix recall (picture & grapheme)	0.078	.272***
Suffix recall (picture only)	0.099	.275***
Determiner and suffix generalisation (consistent)	-0.016	.153*
Determiner and suffix generalisation (inconsistent)	0.096	-0.050

Suffix generalisation (consistent)	0.069	.120*
Suffix generalisation (inconsistent)	0.038	.284***

NB. For procedural memory at test, N= 268, except determiner and suffix generalisation N= 172. For declarative memory at test, N=270, except determiner and suffix generalisation N= 172. Asterisks are defined as: ***p < .001, ** p < .01, *p < .05.

6.3 Chapter Summary and Discussion

The procedural memory task did not show any evidence of sleep-dependent consolidation effects. Consequently, we did not replicate previous findings that sleep leads to improved procedural memory (as in Walker et al., 2002, 2003). The declarative memory task showed evidence of a protective effect of sleep. This supported the decreased post-sleep forgetting previously found in declarative memory using this particular task (as in Rasch et al., 2007).

Given that there *was* evidence of the sleep-related benefit for the declarative task, the prior retrieval of declarative information may have interfered with the retrieval of the procedural information on the session two tests. As discussed in Chapter 1, the idea that interference may occur at retrieval is supported by previous findings. Specifically, there is evidence at the neural level; suggesting that interference arises as a result of the competition between neural areas during both acquisition (Albouy et al., 2015; Brown & Robertson, 2007b; Cohen & Robertson, 2011; Keisler & Shadmehr, 2010; Poldrack et al., 2001) *and* retrieval (e.g. Albouy et al., 2008; Gagné &Cohen, 2016).

Declarative task

We found that participants from both the wake and sleep group showed forgetting of declarative information. Sleep had an additional protective effect against forgetting of declarative information.

Unlike Deliens, Leproult, Neu and Peigneux (2013), we did not find that the protective effect of sleep was driven by SWS. Firstly, and most importantly, this may be because we need a larger sample of participants (for whom polysomnography data was collected, i.e. in Experiment 2) to increase power and potentially see correlations with SWS. Secondly, there may be some key differences between our methodology and Deliens et al.'s (2013) methodology, which may have influenced participants' consolidation. For instance, in their study, the learning criteria (prior to the consolidation delay) were set at 75%. In our study, we did not set learning criteria. The average learning performance was 55.8% for the wake group and 58.3% for the sleep group – this is considerably lower than the learning criteria of 75%. This may be important, given that initial level of acquisition has been previously found to be important for sleep-dependent consolidation. For instance, Tucker and Fishbein (2008) have found that sleep benefits the consolidation of declarative information (in this case word-pair associates learning) only when the initial level of acquisition is high. Hence, SWS may not have benefited our declarative performance, due to a lower initial level of acquisition. Although, given that we found a protective role of sleep relative to wake regarding forgetting, the number of participants (i.e. the statistical power issue) may be more of a problem for the lack of correlations with SWS.

We found a significant test x Experiment interaction in this task as well. Forgetting between learning and retrieval was the greatest in Experiment 6 and the least in Experiment 1. Broadly speaking, it can be suggested that this *may* be because in Experiment 6 the number of stems was increased. In addition, as the additional period of training introduced in Experiment 4 was kept for Experiments 5 and 6, the length of training was also greater in Experiment 6 in comparison to Experiment 1. Therefore, more memory and neuronal resources may have been devoted to language items' consolidation and less to declarative task consolidation as a result. This is something to be explored by collected neural activation data using fMRI in future research.

Finally, it could be suggested that while Deliens et al.'s (2013) word-pair task and our spatial task are both subserved by the declarative system, there may still be crucial differences between these types of language versus spatial tasks. However, we used the same spatial task as Rasch et al., (2007), and did not find a beneficial effect of SWS unlike Rasch et al., (2007). While Rasch et al., (2007) also used both a declarative and a procedural task, potentially resulting in interference; they also included a reactivation paradigm, to enhance the sleep-dependent consolidation process.

Procedural task

Behaviourally, participants from both the wake and the sleep group showed marginally significant forgetting of procedural information, with no additional protective benefit of sleep. Given that at retrieval, the declarative task was tested first, the retrieval of declarative information may have interfered with retrieval of procedural information. This is because the procedural task also has a spatial component (i.e. the spatial layout of the keyboard) that is subserved by the same (hippocampal) system as the spatial information in the declarative task. This idea of the declarative task interfering with the retrieval of the procedural task in session two is supported by previous findings that interference occurs as a result of the *interacting* (competing) nature engagement of the subserving neural areas not only during the acquisition and consolidation of information, but also at retrieval (Albouy et al., 2008; Gagné & Cohen, 2016), as discussed in Chapter 1.

However, this suggestion of interference at retrieval was not tested in *our* Experiments, and hence remains tentative. It would be useful to test participants across the two sessions on the procedural task only, without possibly interfering declarative and language tasks. If when the procedural task is tested *alone* there is indication of sleep-dependent gains (as previously found by Walker et al., 2002), this would confirm that in our Experiments the declarative task may have interfered with the retrieval of the procedural task.

Additionally, there was a marginally significant positive correlation between session two procedural performance and proportion of time in stage 2 sleep. Although unlike Walker et al.'s (2002) study, we did not observe sleep-dependent gains (most likely because of the interference from our other tasks), the same argument that stage 2 sleep plays a role in procedural memory consolidation *might* apply here as well. However, given that this correlation was *marginal*, it remains inconclusive and hence needs to be interpreted with caution. Unlike Fischer et al., (2002), or Wagner et al., (2003), we did not observe any correlation with REM sleep. One possible explanation for this *may* be that stage 2 sleep may play a role when the procedural task is less complex, such as the finger-tapping task, whereas REM sleep may play a role when the procedural task is more complex (Walker et al., 2002).

Correlations with language tasks

We found that performance on session two declarative task retrieval positively correlated with stem recall, suffix recall, and inconsistent suffix endorsement. This suggests that these tasks were all subserved by the same system. The correlation with inconsistent suffix endorsement *may* be indicative of individual suffix knowledge, and not the not the extraction of the suffix-gender regularity extraction (as already suggested in the Results section). This is because the pattern of endorsement on the suffix generalisation which would reflect the extraction of the regularity should have *low* levels of *inconsistent* suffix endorsement. Consequently, *higher* levels of inconsistent suffix endorsement may reflect individual phonological suffix knowledge, *potentially* explaining the positive correlation with the declarative task. Alternatively, this correlation *could* be considered in the context of Ullman's (2001) DP's model predictions about *second* language learning. More specifically, as already mentioned in the Introduction, the DP model (e.g. Ullman, 2001) suggests that in second language learning, the learning of grammatical regularities might initially be related to the declarative time. More time and additional repeated time may be required before grammatical regularities can be linked to the declarative system.

Our findings were in line with previous evidence showing that the consolidation of individual arbitrary word-picture mappings (e.g. Breitenstein et al., 2005), as well as the retrieval of individual trained exemplars (e.g. Lieberman, Chang, Chiao, Bookheimer & Knowlton, 2004) is related to hippocampal activity, which also subserves the declarative system (e.g. Ullman, 2001). Consequently, our correlations suggest that the memory for individual, arbitrary mappings may be linked to the declarative domain-general memory system. This is consistent with the predictions from the declarative-procedural model (Ullman, 2005; Ullman, 2015).

Previous evidence suggests that grammatical regularities may be related to the procedural system (e.g. Brill-Schuetz & Morgan-Short ,2014). Alternatively, grammatical regularities may *initially* be related to the declarative system (via a rapid learning process), but with time (and increased practice) be related to the procedural system (e.g. Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter & Wong, 2014). Despite such evidence, we have not found any correlations with the grammatical regularities, neither in the procedural, nor in the declarative domain, unlike predictions from the declarative/procedural model (e.g. Ullman, 2001). Potentially, as suggested by Ullman

(2001), given that the procedural system requires more exposures for the acquisition and consolidation of regularities, a longer time delay might have been required (than just the 12-hour consolidation delay provided in our Experiment).

Conclusions

Overall, in this exploratory investigation we found no evidence of sleep-dependent consolidation effects for procedural memory. Hence, we did not replicate previous findings that sleep leads to gains in procedural memory (as in Walker et al., 2002, 2003). The declarative memory task showed evidence of a protective effect of sleep, in line with previous findings using the same declarative task (Rasch et al., 2007) that forgetting is decreased followed sleep. In addition, we found correlations between arbitrary measures and the declarative task, in line with previous findings (e.g. Breitenstein et al., 2005) that the consolidation of individual arbitrary mappings may be linked to the declarative domain-general memory system, as predicted by the declarative-procedural model (Ullman, 2005; Ullman, 2015).

Chapter 7: General Discussion

The six Experiments in this thesis investigated the role of sleep in the consolidation of individual, arbitrary mappings, and in the extraction of regularities contained within systematic mappings, based on the predictions of the CLS model (e.g. McClelland et al., 1995). We also addressed the factors in the language input that may affect the extraction of regularities regardless of sleep. In Experiments 1 and 2 we investigated whether there is extraction of fully systematic determiners, and suffixes at the intermediate level of systematicity. In Experiments 3 and 4 we investigated whether the temporal order of information affects extraction (i.e. determiners being at the beginning of the word are encountered earlier in time than suffixes) by swapping the systematicity of the two language mappings. Also in Experiment 4 we investigated whether additional training boosts extraction, as well as whether extracted regularities are available immediately but forgotten following a consolidation delay. In Experiments 5 and 6 we investigated whether there is a redundancy of multiples cues to gender by removing determiners from the input. Also in Experiment 6 we investigated whether increasing stem variability would facilitate suffix-gender regularity extraction. Firstly, the key findings are summarised in terms of the language input factors regardless of sleep.

7.1 Thesis Summary

7.1.1 Chapters 2 and 3

In Experiments 1 and 2 (Chapters 2 and 3) we investigated the distinction between arbitrary and systematic mappings and also tested the effect of varying the degree of systematicity. We used a novel language which implemented arbitrary mappings at the level of individual words (arbitrary form-meaning mappings), and systematic mappings in the regularities across a set of words (implemented as determiners and suffixes mapping onto the natural gender of the referents). We also varied the degree of the systematicity of the mappings. Specifically, in Experiments 1 (Chapter 2) and 2 (Chapter 3), determiners were fully systematic (1 determiner per gender), while suffixes were of intermediate systematicity (2 suffixes per gender).

In Experiment 1 and 2, there was evidence of determiner-gender regularity extraction, i.e. extraction of the fully systematic mapping. While findings for suffixes were mixed, there was no strong evidence of suffix-gender regularity extraction, i.e. no extraction of the mapping at the intermediate level of systematicity. This was not in line with previous findings that both infants and adults are able to extract statistical regularities from *phonological* sequences (e.g. Gomez et al., 2006; Saffran, Aslin & Newport, 2006; St Clair, Monaghan & Ramscar, 2009). Crucially, such studies have not used semantic cues. This is important, because the structure of the language input changes the nature of learning and consolidation (e.g. Ramscar et al., 2010). Our language input differed (i.e. was more complex) to that provided by previous studies showing extraction of statistical regularities.

Specifically, given that we had two phonological cues of varying systematicity (determiners and suffixes) mapping on to the *semantic* mapping of gender, we provided participants with a complex language input. To elaborate, because of the nature of the training language, participants were exposed to sound-meaning mappings of varying systematicity (fully systematic determiner-gender mapping, less systematic suffix-gender mapping). Participants were *also* exposed to individual stem-picture mappings, and morphological determiner-suffix co-occurrence mappings. Semantically, participants were exposed to the shared semantic regularity of the gender mapping (i.e. all pictures were of human occupations of either male or female gender). Phonologically, participants could acquire separately individual determiners, individual stems and individual suffixes.

All of the above illustrates that the language input was complex, and there is already evidence to suggest that gender is difficult for adult second language learners to acquire, even with repeated practice (e.g. Dewaele & Véronique, 2001; Scherag, Demuth, Rösler, Neville & Röder , 2004). In addition, having so many different mappings in the input (in particular with the inclusion of semantics) competing for the learner's attention makes the learning "competitive". Given the limits on the learner's attention, the learner needs to select the cues and mappings which will be of the highest use in future language use. Hence, competition for the highest predictive value arises between cues and mappings (Ramscar, Yarlett, Dye, Denny & Thorpe, 2010; Ramscar, 2013). This is not the case with statistical learning literature where for instance, only *phonological* cues are provided, but not semantic cues (e.g. Gomez et al., 2006).

Given that we did not see strong evidence of suffix-gender regularity extraction in Experiments 1 and 2, it may be suggested that having multiples cues and mappings in the language input makes learning more competitive, and hence changes the learning process. Consequently, on the basis of the findings of Experiments 1 and 2, it was hypothesised that a potential reason we did not see evidence of suffix-gender regularity extraction was because of the suffixes' lower degree of systematicity, in comparison to determiners.

7.1.2 Chapter 4

Having shown that in competitive learning (given the nature of our training language), fully systematic determiner-gender regularities are extracted, but suffix-gender regularities of intermediate systematicity are not, it was of interest whether this finding can be explained purely on the basis of systematicity. We investigated this by swapping the systematicity of the mappings in Experiments 3 and 4 (in comparison to the mappings in Experiments 1 and 2). We also provided additional training in Experiment 4 (making the initial memory traces stronger). This was done to investigate whether differences in wake-dependent versus sleep-dependent consolidation for the extraction of regularities would emerge as a result of the additional training. Also, as an additional question in Experiment 4 we tested an additional immediate group to see whether the regularities could be extracted immediately, but lost after a delay.

As a reminder, whereas determiners were fully systematic in Experiments 1 and 2, we made them of intermediate systematicity in Experiments 3 and 4. Whereas suffixes were of intermediate systematicity in Experiments 1 and 2, we made them fully systematic in Experiments 3 and 4. Given the nature of our training language, the more systematic cue (i.e. the more reliable cue) was also the more available (i.e. more frequent) cue in the input. The concepts of systematicity (i.e. reliability) and frequency (i.e. availability) have been discussed within the Competition model (e.g. Bates & MacWhinney, 1989). As already discussed in the previous Chapters, this model suggests that cues with higher reliability and availability are judged to have higher predictive value, and hence are more likely to be acquired than cues with lower reliability and availability. On the basis of this model, and also because the fully systematic determiner-gender regularities were extracted in Experiments 1 and 2, we predicted that the fully systematic suffixes should be extracted in Experiments 3 and 4.

Moreover, the ideas from Rescorla and Wagner's (1972) model about mechanisms of learning are also applicable to our stimuli. They have suggested that learning about the meaning, i.e. "outcome" of a cue involves processing not only the consequences of that specific cue being *present* on trials. Learning about the meaning, i.e. "outcome" of a cue also occurs on trials where that specific cue is *absent*. This argument can be applied to our language stimuli, with the determiners and suffixes as the predictive cues, and the semantic gender (of the picture) being the "outcome" to be predicted. The above suggests that participants learn not only how often the determiners and suffixes co-occur with the corresponding gender mapping, but also how often they do not. To elaborate, when suffixes were the intermediate systematicity mapping in Experiments 1 and 2, there were 2 suffixes per gender. This means that gender was associated with 2 suffix cues. Consequently, knowing the gender of the pictured character did *not* help in knowing which of the two suffixes will be used.

This means that learning occurred not only every time a feminine character was seen paired with the suffix *'eem'*, but also every time that a feminine character was seen but the suffix *'eem'* was absent. This made the intermediate systematicity cue less predictive of gender. This *may* explain why there was no strong evidence of suffix-gender regularity extraction in Experiments 1 and 2. Consequently, if systematicity is what is important for regularity mapping acquisition, then we should see extraction of the suffix-gender regularities in Experiments 3 and 4, given that fully systematic determiners were extracted in Experiments 1 and 2.

For determiner-gender regularities (now the mapping at the intermediate level of systematicity), both Experiments 3 and 4 provided evidence to suggest that individual determiner-picture mapping memory was low. In Experiment 3 there was *no* evidence of determiner-gender regularity extraction in either of the groups. In Experiment 4 when additional training was provided, there was evidence to suggest that determiner-gender regularities were extracted only immediately (but not after a delay, albeit with some preservation of extraction following wake-dependent consolidation only).

For suffix-gender regularities (now the fully systematic mapping), the findings were inconclusive in Experiment 3. In Experiment 4, when additional training was provided, there was no strong evidence of suffix-gender regularity extraction. There was however *some* evidence that suffix-gender regularity emerged following wake-dependent consolidation only. There was no evidence of suffix-gender regularity extraction immediately, despite the full systematicity of the mapping. These findings suggest that in language, systematicity is important, but it is not all about systematicity.

In addition, as within our studies we focused on consolidation effects (i.e. the differences in consolidation of individual, arbitrary mappings, in comparison to the consolidation of regularities contained within systematic mappings), we did not investigate the predictive relationships within the language. It is interesting to consider the cue and stimulus both being *within* the language, i.e. the predictive relationship between the determiners and the suffixes (as opposed to the determiner-picture gender and suffix-picture gender regularities).

In Experiments 3 and 4, there were two determiners per suffix. This means that seeing the determiners *'tib'* or *jov'* at the beginning of the word would reliably cue the suffix *'eem'*. Seeing the determiners *'ked'* or *'paz'* would reliably predict the suffix *'ool'.* The determiners were encountered first (i.e. appeared prior to suffixes temporally), due to being at the beginning of the word. Hence, participants could have used the determiners as a cue to figure out the correct suffix, based on the probability of the determiner-suffix co-occurrence. This means that participants did not *need* to extract the suffix-gender regularity, given that the determiners cued both the *same* gender regularity, *and* the upcoming suffix.

Whether participants figured out the transitional determiner-suffix probability or not, is an interesting point to consider in further research. This could be done by including an additional grammatical judgment *generalisation* task, using previously unseen items (i.e. new stems paired with trained determiners and suffixes). The picture characters depicting gender could be removed from this task, so that the participants could not rely on the gender of the referents to figure out the determiners and suffixes. As with our already existing generalisation tasks, half of the items would be consistent (e.g. a word whereby the determiner *'tib'* is paired with the suffix *'eem'*), and half would be inconsistent (e.g. a word whereby the determiner *'tib'* is paired with the suffix *'ool'*). Participants should be able to discriminate between the consistent and inconsistent items, and endorse the consistent items as grammatically correct more than the inconsistent items, if they had acquired the correct determiner-suffix probability. It is expected that such a task would indeed show that participants had learned the determiner-suffix co-occurrence, as this would be in line with the existing findings from the statistical literature that both infants and adults are able to learn transitional probabilities within language, for instance using non-adjacent dependencies (see Romberg and Saffran, 2010 for a review of statistical learning literature).

The expected finding that participants were also learning the transitional determiner-suffix probabilities, would also be consistent with the predictions of Rescorla and Wagner's (1972) model, whereby the determiner would be treated as the cue, and the word (and the suffix) would be treated as the stimulus. On the basis of this model, it could be suggested that there is an associative strength between the determiner and the suffix. As already described earlier in relation to language-picture mappings, this strength would be determined on the basis of not only each time that a determiner cue was *present*, but also each time that a determiner cue was *absent*. Specifically, participants would learn about the predictive relationship between the determiners and suffixes not only each time that both the determiner *'tib'* and the suffix *'eem'* occurred, but also each time that the determiner *'tib'* was *absent*, but the suffix *'eem'* occurred.

Finally, as already discussed, previous findings have shown that both infants and adults can extract statistical regularities from purely *phonological* input (e.g. Gomez et al., 2006; Saffran, Aslin & Newport, 1996; St Clair, Monaghan & Ramscar, 2009). However, as already suggested, the majority of such studies have not addressed competitive learning, i.e. what happens when the learner is faced with multiple cues in the input (cues which also include semantics). Meanwhile, models such as the Competition model (e.g. Bates & MacWhinney, 1989) and Rescorla and Wagner's model (1972) *have* discussed the importance of systematicity in the case of competitive learning. Indeed, in Experiments 1 and 2, when determiners were the more systematic cue and suffixes were the less systematic cue, our findings supported the predictions of the Competition model (e.g. Bates & MacWhinney, 1989) and Rescorla and Wagner's (1972) model. Yet, when the systematicity was reversed and determiners became the less systematic cue, and suffixes the more systematic cue, our findings no longer supported the predictions of the models, given that we did not see strong evidence of suffix-gender regularity extraction, and yet we saw evidence of determiner-gender regularity extraction immediately. This suggests that the ordering of information in time is crucial.

While the Competition model (e.g. Bates & MacWhinney, 1989) and Rescorla and Wagner's model (1972) have addressed the issue of competitive learning, the importance of temporal ordering of information has not been addressed. More specifically, we are referring to the temporal importance in terms of the word order, i.e. the encountering of information in time. Specifically, in our Experiments, determiners are encountered earlier in time than suffixes, due to word order: determiners are at the beginning of the word, whilst suffixes are at the end. Potentially, as a result of this temporal difference, determiners may be differentially treated than suffixes. This will be discussed further in the next subsection in the context of the findings from Chapter 5.

7.1.3 Chapter 5

Our lack of strong evidence for suffix-gender regularity extraction in Experiments 1-4 contradicts previous evidence pointing to the suffixing preference, i.e. that suffixes as a cue to grammatical class are acquired better than prefixes (e.g. St Clair, Monaghan & Ramscar, 2009). Whilst we have been using determiners as opposed to prefixes, it is reasonable to suggest that the same conclusions may apply. Yet, we found evidence of determiner-gender regularity extraction, but no strong evidence of suffix-gender regularity extraction, even when suffixes were fully systematic.

These findings may be explained by the concept of redundancy as a consequence of multiple cues to the same function (e.g. DeKeyser, 2005), as already suggested in Chapter 5 itself. Given that both determiners and suffixes predicted the same gender mapping, the presence of determiners may have made the suffix cue redundant. Conjecturally, as already suggested, one reason why determiners were preferred over suffixes may be because of the temporal ordering of information in time, i.e. that determiners are positioned at the beginning of the word.

To elaborate on this (and as already discussed in Chapter 5 itself), it is important to bear in mind that our participants are adult native English speakers. Based on corpus analyses of English, Monaghan, Christiansen and Fitneva (2011) have shown that information at the beginning of the word is typically processed for individual meaning, i.e. the identity of the word. Meanwhile, information at the end of the word is typically processed in order to systematically categorize the word. These findings can be related to Van Patten's input processing model (e.g. VanPatten 2004), which suggests that participants' priority is in the acquisition of the individual meanings of words, prior to the acquisition of grammatical categories. Consequently, based on our participants' previous experience with the acquisition of English as a first language (see Ellis, 2006 for a discussion of the importance of L1 in second language learning), our participants may have prioritized the processing of the beginnings of words (i.e. the determiners) based on their “expectancy” to receive word identity information, over processing the suffixes. In line with this, according to the processing heuristic (MacWhinney, 1999), learners pay more attention to information at the beginning of a word, than to information at the end of the word. Indeed, Frigo and McDonald (1998) have shown that in the acquisition of gender regularities, participants relied on the beginning marker regardless of salience, but on the end marker only when it was more salient.

Secondly, it is important to bear in mind the perceivability of cues. As shown by MacWhinney et al., (1985) in Hungarian, even highly reliable cues may not be acquired if they are not easy to identify, i.e. have low perceivability. Consequently, cues with lower perceivability are acquired later than more perceivable cues. Given that in our stimuli in Experiments 1-4 determiners were separate from the word, they may have been a more perceivable cue to gender than suffixes.

To summarize, determiners (due to their position at the beginning of the word) are encountered earlier in time than suffixes. Native English speakers' attention may be drawn to the beginnings of words, as a result of their priority for meaning acquisition (over grammatical category acquisition). Consequently, determiners may receive more processing than suffixes. Secondly, determiners due to being separate from the word (unlike suffixes), may have higher perceivability than suffixes. These are potential factors which may result in participants' preference for determiners over suffixes. Consequently, the determiner cue may be assigned higher cue strength in predicting gender. Given that both determiners and suffixes are cues to the same gender, suffixes may be deemed redundant in the presence of the higher cue strength determiners. Consequently, the presence of the determiner-gender regularity may inhibit suffix-gender regularity extraction.

In order to investigate whether the presence of determiner-gender regularities in the input may have indeed been inhibiting suffix-gender regularity extraction (in our Experiments 1-4), in Experiments 5 and 6 we removed determiners from the input (keeping the additional level of training introduced in Experiment 4). In keeping with the predictions of the Competition model (e.g. Bates and MacWhinney, 1989), we kept the suffix mapping fully systematic, in order to strengthen the validity of the suffix cue in predicting gender. Despite this, we did not see strong evidence of suffix-gender regularity extraction. Making suffixes fully systematic and removing determiners from the input was *not* sufficient to result in strong suffix-gender regularity extraction (albeit *some* indication of extraction in the wake group only, driven by better training performance).

Consequently, in Experiment 6 we investigated whether increased stem variability would boost suffix-gender regularity extraction in isolation of the determiners in the input. As already discussed in Chapter 5 itself, increasing exemplar variability has previously been found to benefit the extraction of regularities in the phonological domain using an artificial language (e.g. Gomez, 2002; Onnis, Christiansen, Chater & Gomez, 2003; Onnis, Monaghan, Christiansen & Chater, 2004). Additionally, increased stem variability has been found to benefit the extraction of suffix-gender regularities in Russian, i.e. in natural language (Eidsvag, Austad, Plante and Asbjornsen, 2015). The above evidence has not tested extraction following a consolidation delay. Tamminen et al., (2015) tested the effect of increased exemplar variability following a one week consolidation delay using already known stems but novel affixes. They have shown that increasing the number of stems per affix facilitated affix generalisation.

In our Experiment 6, we were interested in the effect of increased stem variability following a 12-hour consolidation delay, using our artificial training language (as opposed to known stems such as Tamminen et al., 2015, or a fully natural language such as Eidsvag et al., 2015). We utilized phonology-meaning mappings, as opposed to phonology alone (such as Gomez, 2002; Onnis, Christiansen, Chater & Gomez, 2003; Onnis, Monaghan, Christiansen & Chater, 2004).

As expected, when stem variability was increased in Experiment 6, there was evidence that the suffix-gender regularity was extracted, regardless of sleep. This may be because in line with Gomez's (2002) explanation, increased exemplar variability (which in our case was accomplished by increasing the number of stems) may boost the salience and the detectability of the regularity (in our case the suffix cue). This is in line with suggestions from the Competition model (e.g. Bates & MacWhinney, 1989) that increased perceivability of a cue strengthens the validity of that cue, making it more likely to be acquired.

Interpreted within CLS models (e.g. McClelland et al., 1995), the consequence of suffixes being seen co-occurring with a larger number of stems, is a greater number of lexical representations which share the same suffix-gender regularity. Consequently, the suffix overlap in the neocortex is greater, boosting the parsing of the suffixes from the stem, and in turn leading to extraction.

In addition, based on Ford et al., (2010), and Monaghan et al., (2005), *frequency* may be the factor that boosts suffix-gender regularity extraction. To elaborate on this point, increasing the number of stems per suffix meant that suffixes were used across a larger number of exemplars. Consequently, increasing stem variability in turn increased the frequency of *use* for suffixes. This view that increasing stem variability increased frequency of suffix use (per stem) may also be informative in light of usage based grammar theories. In line with such theories, language acquisition (and associative learning of language) is exemplar-driven, and consequently frequency of usage is a key factor in language acquisition (e.g. Ellis 2002; 2006, Ellis & Larsen-Freeman, 2009). Participants parse out regularities from their experience with exemplars in the input. In our Experiment 6, the consequence of increased stem variability was a greater number of exemplars in the input which were consistent with the suffix-gender regularity. This may have boosted the construction of the suffix-gender regularity schema, which would have boosted extraction and generalization.

To elaborate, within the usage-based framework and construction grammar, form-meaning mappings are termed constructions (e.g. Tomasello, 2003). Morphological regularities (such as those contained within our determiner and suffix mappings) map on to semantic regularities (such as our gender). Thus, a schema of constructions, i.e. form-meaning mappings is created, from which regularities can be extracted. Given that learning is driven by the available exemplars, frequency of use (i.e. how many stems a suffix is seen paired with in the input) is crucial for learning. Each time a suffix is seen in the input with a different stem, the strength of the memory representation of the form-meaning mapping (such as the suffix-gender "construction") is increased.

Moreover, it is important to bear in mind that the learner's goal is to reduce future uncertainty (e.g. Ramscar et al., 2010), i.e. to increase future predictability of gender/meaning in using the extracted regularities such as determiners and suffixes. A greater number of exemplars with the same shared suffix-gender regularity increases the certainty that the suffix-gender regularity *does* indeed reliably predict gender *generally*, rather than in just a small subset of items (Bybee & Thompson, 2000). Consequently, regularity acquisition is more robust when the number of exemplars is increased.

Having said the above, it is important to mention an alternative explanation. In Experiment 6 we found evidence to suggest that when stem variability was increased, although stem consolidation was not affected, suffix recall was decreased. Increased stem variability potentially resulted in lower individual suffix memory. Thus, suffix-gender regularity extraction *may* have been boosted as a consequence of lower individual suffix memory. This suggestion is supported by previous findings (e.g. Vlach et al., 2012, Werchan & Gomez 2014) that the consolidation of individual elements may inhibit generalization. Consequently, some forgetting of individual items may be needed in order to boost extraction and generalization.

7.1.4 Summary of language findings

The learners' task in language acquisition is to reduce future uncertainty, or increase future predictability (e.g. Ramscar, 2010). For instance, in our case this is true in using the phonological systematic mappings (determiners and suffixes) to predict the natural gender of the word.

As shown in Table 17, our findings suggest that it is an oversimplification to think that language learning depends purely on "probabilities", i.e. the degree of systematicity or purely statistical predictability of cues. This view applies in the case of noncompetitive learning, such as the acquisition of phonology alone, such as has been used in the majority of statistical learning literature (e.g. Gomez et al., 2006; also see Romberg and Saffran (2010) for a review of statistical learning literature). Yet, when semantics are added to the input, the acquisition process becomes more competitive, consequently changing the nature of learning and consolidation.

In our Experiments, determiners and suffixes were not found to have equal predictive value on the basis of systematicity alone. We found evidence of determiner-gender regularity extraction following a consolidation delay when determiners were fully systematic. We also saw evidence of determiner-gender regularity extraction *immediately* when determiners were of intermediate systematicity. In contrast, we did not see strong evidence of suffix-gender extraction when suffixes were fully systematic, even when determiners were removed from the input. In order for suffix-gender regularity extraction to emerge, increased exemplar variability (i.e. increased stem variability) was also needed. This suggests that increased number of exemplars with the same shared regularity may increase the learner's certainty that the regularity is a reliable predictor of gender, and may strengthen the created suffix-gender regularity representations.

Table 17. Results of systematic mappings (regardless of sleep) across 6 Experiments.

	Suffixes Extracted?	Determiners extracted?
Experiment 1: Determiners = full systematicity Suffixes = intermediate systematicity	No strong evidence. Against chance: Sleep – no evidence of extraction Wake – evidence of extraction	Good individual determiner-picture mapping memory (determiner selection task) Evidence of determiner-gender regularity extraction in both groups
Experiment 2: Replication of Experiment 1 with polysomnography setup	No evidence of extraction	Yes
Experiment 3: Determiners = intermediate systematicity Suffixes = full systematicity	Inconclusive, as the pattern of extraction was the opposite to what was expected. Overall low endorsement: sleep group at chance for both, wake group above chance for inconsistent only	Low memory, determiner error analysis showed that participants relied on individual mappings, not regularities (at chance) Low endorsement, at chance on all items, no evidence of extraction
Experiment 4: Replication of Experiment 3 with increased training, immediate group	No strong evidence of extraction (immediately/delay) Against chance:	Low memory, Error analysis: at chance (relied on individual mappings, not regularities).

included	Wake group only began to extract (not immediate or sleep)	Higher endorsement than in E3, no longer close to chance.

Table continued 17 continued.

	Suffixes Extracted?	Determiners extracted?
Experiment 4 continued		Evidence of extraction in the immediate group (both main effect of consistency, and against chance performance) No main effect of consistency in the delayed groups. Against chance performance of the delayed groups: Sleep group: above chance on all items, no evidence of extraction. Wake group: evidence of extraction
Experiment 5: no determiners in the input, suffixes were fully systematic	No strong evidence of extraction. Wake group: evidence of extraction (a marginally significant three-way interaction between task order, consistency and group)	n/a

Table 17 continued.

	Suffixes Extracted?	Determiners extracted?
Experiment 5 continued	Against chance: wake = evidence of extraction Sleep = no evidence of extraction	
Experiment 6: No determiners in the input, suffixes were fully systematic, stem variability was increased	Yes, in both groups	n/a

7.1.5 Summary of language version differences on the training task

In all of our Experiments, all of the language tasks were counterbalanced on the language learnt (in reference to the determiner-natural picture gender mappings). We were not expecting to find any differences between language versions at training. However, we found *some* indication across the six Experiments that some differences between language versions *may* have emerged. To account for this, any differences according to language version found on the training tasks across the six Experiments are summarized below, in order to see whether a typical pattern emerged (e.g. whether language version one was typically better acquired than language version two).

As a reminder (as has already been described in Chapter 2 Method), in Experiments 1 and 2, for half of the participants the determiner '*tib'* and the suffixes '-*eem'* and '-*esh'* mapped onto pictures of human occupations whose natural gender was female and the determiner '*ked*' and the suffixes *'-ool'* and *'-aff'* mapped on to male and vice versa for the other half of the participants (i.e. '*tib'* = male, '*ked'* = female). In Experiments 3 and 4, the determiners *'tib'* and *jov'* and the suffix *'eem'* mapped onto one gender. The determiners *'ked'* and *'paz'* and the suffix *'ool'* mapped onto a different gender.

In Experiment 1, the analysis of the correct target RTs revealed a marginally significant interaction between group and language. Participants in the wake group improved more in language version 1, but participants in the sleep group improved more in language version 2. There were no other indications of differences on the basis of the language version in Experiment 1. There were no indications of differences on the basis of the language version in Experiment 2.

In Experiment 3, the analysis of correct target RTs revealed a marginally significant interaction between block and language. Regardless of group, improvement across blocks was greater in language version 1, than in language version 2. There were no other indications of differences on the basis of the language version in Experiment 3. In Experiment 4, there were no indications of differences on the basis of the language version.

In Experiment 5, the analysis of the correct target RTs revealed a marginally significant main effect of language version, and a significant interaction between group and language version. Namely, the wake group had less improvement in target RTs on language version 1, but more improvement in target RTs on language version 2. There were no other indications of differences in the basis of the language version in Experiment 5.

In Experiment 6, the analysis of the correct target RTs revealed that participants in language version 2 had greater improvement across the blocks than participants in language version 1. However, the analysis of the correct noise RTs revealed the opposite (although marginal) result. Namely, participants in language version 1 had marginally greater improvement across the blocks than participants in language version 2.

Consequently, across Experiments, differences which emerged on the basis of language version were firstly typically found on the correct target RTs' analyses (and not on the accuracy analyses). Secondly, aside from Experiment 5 these differences were marginal, and hence strong conclusions cannot be made. No obvious patterns emerged in terms of language version, i.e. it was *not* that one language version typically showed a larger improvement than another regardless of group. No patterns in differences between groups according to language have emerged.

Indeed, no such differences between language versions were expected. While this would need further research, at this stage it can be concluded that no strong patterns regarding differences in the learning of language versions were found across the six Experiments.

7.1.6 Chapter 6

In Chapter 6 we firstly sought to link aspects of the language tasks with the declarative/ procedural model (e.g. Ullman, 2005) in an explorative investigation. We found correlations between individual, arbitrary mappings (namely the individual stems and suffixes) with the declarative task, as predicted by the model. This was also in line with previous evidence suggesting links between the learning of individual, arbitrary mappings and the declarative system, in line with suggestions that both are subserved by the hippocampus (e.g. Breitenstein et al. 2005; Ullman, 2001). We did not see any correlations with the systematic mappings and the procedural task. This was not in line with previous evidence linking the retrieval of grammatical regularities and the procedural system, given suggestions that both are subserved by the basal ganglia (including the caudate nucleus) (e.g. Lieberman et al., 2004; Ullman, 2001).

Secondly, in Chapter 6 we also sought to explore whether procedural and declarative memories are affected differently by sleep-dependent consolidation compared to wake-dependent consolidation (i.e. whether sleep benefits the consolidation of procedural and declarative information). For procedural memory, we did not replicate the post-sleep improvement in speed (previously found by Walker et al., 2002; 2003). Performance was largely maintained (with marginal loss of speed) regardless of the consolidation delay type, i.e. regardless of sleep or wake. Given the lack of behavioural effect of sleep, we did not expect to find any correlations with sleep parameters for procedural memory. Indeed, unlike Fisher et al., (2002) and Wagner et al., (2003), we did not find correlations between change in procedural memory (between training and retest) and the proportion of time spent in REM sleep. We found a marginal correlation between change in procedural memory (between training and retest) and the proportion of time spent in stage 2 sleep (previously found by Walker et al., (2002). However, as this correlation was only marginal, this remains inconclusive and further research is needed to clarify the existence of this correlation.

For declarative memory, we found forgetting between training and retest, with sleep having a protective effect against forgetting. This was in line with previous findings using the same declarative task (Rasch et al., 2007). However, unlike Rasch et al. (2007) we did not find this sleep-dependent preservation to be related to SWS. We suggested that potentially a higher initial level of learning would have been needed in order for sleep (and SWS in particular) to have more an effect on declarative memory (e.g. Tucker & Fishbein, 2008).

Overall, given that Chapter 6 was an exploratory investigation, further research is still needed to clarify the findings.

7.1.7 Sleep and arbitrary/systematic language mappings

Across the six Experiments in this thesis we addressed a new dimension in the learning and consolidation of sound-meaning mappings. Namely, we addressed the difference in learning and consolidating of arbitrary versus systematic mappings, based on the connectionist framework. More specifically, according to the Complementary Learning Systems model (e.g. McClelland, McNaughton & O'Reilly 1995) the learning of arbitrary mappings should be more influenced by sleep-associated consolidation relative to the learning of systematic mappings. This is because arbitrary mappings rely more on the hippocampal systems, whereas systematic mappings rely more on neocortical systems. Arbitrary mappings are learned as individual items, whereas shared regularities which can be applied to previously unseen exemplars are extracted from systematic mappings.

We were interested in the role of sleep in the consolidation of arbitrary mappings, and the extraction of regularities. According to two-stage systems consolidation models (e.g. O'Reilly & Norman, 2002; Rasch & Born, 2007), sleep mediates the hippocampal-neocortex information dialogue, thus facilitating the integration of newly acquired material into already existing schemas. Hence, sleep should benefit information which is more hippocampally-dependent. Indeed, there is some evidence to suggest that sleep does benefit the consolidation of mappings which are more hippocampally-dependent (e.g. Rasch et al., 2007). Given that arbitrary mappings are more hippocampally-dependent, they should rely on the hippocampal-neocortical dialogue more than systematic mappings.

Consequently, on the basis of the connectionist framework, both the CLS (e.g. McClelland et al., 1995) and the two-stage systems consolidation models (e.g. O'Reilly & Norman, 2002; Rasch & Born, 2007) we predicted a benefit of sleep for the consolidation of arbitrary mappings, but not for the extraction of regularities. It was important to test these predictions, given that as discussed in Chapter 1, there is evidence in the previous literature which suggests that sleep *does* benefit the extraction of regularities (e.g. Gomez et al., 2006), in contrast to our predictions.

In Experiments 1 and 2, we found a benefit of sleep for the consolidation of individual mappings at the arbitrary level (individual stems and suffixes), as measured by a recall task, consistent with our predictions on the basis of the CLS (e.g. McClelland et al., 1995). To analyse suffix abstraction we looked at the suffix errors made on the recall task. We found some evidence that the sleep group, but not the wake group, produced correct suffix-gender mappings above chance when making errors for the picture only cue type. This provided some evidence to suggest that sleep may have facilitated the abstraction of the suffix-gender regularity when suffixes were at the intermediate level of systematicity. However, the two generalisation tasks did not provide any additional evidence for a benefit of sleep for extraction of regularities. Rather, the performance of the wake group was more in line with suffix-gender regularity extraction on the suffix generalisation task, than performance of the sleep group. Determiner-gender regularities were extracted regardless of the consolidation delay type (i.e. regardless of whether the consolidation delay included a period of daytime wakefulness or overnight sleep).

In addition, in Experiment 2 (Chapter 3) we also collected overnight polysomnography data to correlate the language measures with sleep parameters. We found a negative correlation between SWS and stem recall, i.e. the individual arbitrary mapping. Although not in line with our predictions, this correlation was in line with Payne et al.'s (2009) negative correlation between SWS and veridical trained recall using a DRM paradigm.

While we did not see evidence of sleep for the extraction of regularities, unlike previous studies (e.g. Gomez et al., 2006, using phonology alone), the argument that the nature of the language input changes the nature of learning and consolidation (e.g. Ramscar et al., 2010) might also be extended to sleep-dependent consolidation. For instance, studies using cross-modal input (e.g. Hennies, Lewis, Durrant, Cousins & Ralph, 2014; Sweegers & Talamini, 2014) have provided evidence for the benefit of wake, as opposed to sleep for the extraction of regularities.

In Experiments 3 and 4, we swapped the systematicity of the mappings. We also provided additional training in Experiment 4 to make the initial memory traces stronger.

Also in Experiment 4 we tested an immediate group to see whether the regularities could be extracted immediately, but lost after a delay. For the arbitrary mappings, there was no statistical evidence that sleep benefited stem recall. There was statistical evidence in Experiment 3 (with less training) that sleep benefited individual suffix recall, but this statistical benefit was not found in Experiment 4 (when more training was provided). Individual determiner-picture mapping memory was low, and there was no evidence of determiner-gender regularity extraction in either of the delayed groups in Experiment 3 (when less training was provided). In Experiment 4, when additional training was provided, determiner-gender regularities were extracted immediately. There was *some* indication (based on above chance endorsement of consistent mappings only) that the wake group, but not the sleep group preserved some of the determiner-gender regularity extraction from being forgotten.

For suffixes, findings from Experiment 3 were inconclusive given that the pattern of endorsement was the opposite of that expected for extraction. In Experiment 4 (with additional training), there was no *strong* evidence of suffix-gender regularity extraction, neither immediately, nor after a delay. There was however *some* indication (again, based on above chance performance on consistent mappings only, but not inconsistent mappings on the suffix generalisation task) that the wake group only, (but not the sleep or immediate group) had *started* to extract the suffix-gender regularities.

In Experiments 5 and 6 we removed determiners from the input, and additionally in Experiment 6 we increased stem variability. For the arbitrary mappings, there was no statistical evidence for the benefit of sleep over wake in consolidating individual stems and suffixes in Experiment 5. This benefit of sleep for consolidating individual stems and suffixes was statistically significant in Experiment 6 (when stem variability was increased). For the systematic mappings, in Experiment 5 (when stem variability was lower) there was evidence to suggest that the performance of the wake group was more in line with suffix-gender regularity extraction, than the performance of the sleep group. However, there was indication that this evidence of extraction in the wake group, was *not* a result of wake-dependent *consolidation*, but rather due to better initial levels of acquisition. In Experiment 6, when stem variability was increased, both groups extracted the suffix-gender regularities, with no additional benefit of sleep.

The above evidence suggests that sleep may be beneficial for consolidating individual, arbitrary mappings (as shown in Table 18), but not for the extraction of regularities. Crucially, we are *not* suggesting that sleep cannot or does not under any circumstances benefit regularity extraction. Indeed, as discussed in Chapter 1, previous studies *have* found a benefit of sleep for regularity extraction (e.g. Gomez et al., 2006). Rather, we are suggesting that the nature of learning and also consolidation may change when the learning input and language structure becomes more complex (e.g. Ramscar et al., 2010). As already discussed in the above subsection of the General Discussion (e.g. Chapters 2 and 3 subsection), we are suggesting that we had a relatively greater number of mappings in our input (e.g. determiner-suffix co-occurrences, individual determiner-picture mappings, individual stem-picture mappings, determiner-gender regularities, suffix-gender regularities). This may have made our input relatively more complex and hence made the learning relatively more competitive. Consequently, the nature of consolidation may have also been altered (in comparison to a relatively less complex input).

In the presence of a complex input with a relatively greater number of mappings, sleep may have preferentially benefited the consolidation of individual arbitrary mappings, over systematic mappings. The findings are consistent with the CLS (e.g. McClelland et al., 1995), based on which it might be suggested that systematic mappings rely more on the neocortical system, whereas arbitrary mappings rely more on the hippocampal system. Given that the benefit of sleep is in mediating the hippocampal-neocortical dialogue (e.g. Rasch et al., 2007), in a complex input with many different mappings, sleep may preferentially benefit the mappings which are more hippocampally-dependent, such as our arbitrary mappings, over the less hippocampally-dependent systematic mappings.

Table 18. Results of the effect of sleep in consolidating individual, arbitrary mappings across 6 Experiments.

Experiment (E)	Individual Stems: sleep benefit?	Individual Suffixes: sleep benefit?
E1	Yes	Yes.
E2	Yes, Negative correlation with SWS%	No correlation with sleep parameters
E3	Numerically, but not statistically	Yes
E4	Numerically, but not statistically	Numerically, but not statistically
E5	Numerically, but not statistically	Numerically, but not statistically
E6	Yes	Yes

For systematic mappings, although across the six Experiments of this thesis the evidence for the wake benefit for extraction was not strong, it may be important to discuss the *possibility* of wake (but not sleep) boosting extraction. Specifically, in Experiment 4, there was *some* indication of extraction of both the determiner-gender regularities and the suffix-gender regularities by the wake group, but not by the sleep group. Similarly, in Experiment 5 when no determiners were present in the input, there was indication of extraction of suffix-gender regularities in the wake group, but not in the sleep group. It could be suggested that the initial levels of acquisition might have played a role in both of these Experiments.

Specifically, in Experiment 4 the correct target RTs analysis revealed that the wake group improved more than the sleep group, reflective of greater learning (the greater learning of the wake group in comparison to the sleep group was also supported by the accuracy analysis at the end of training). Having said this, when session 1 training performance was included as a covariate in the session two analyses, there was *no* indication that training performance affected determiner-gender regularity extraction or suffix-gender regularity extraction in Experiment 4.

In Experiment 5, the wake group had higher performance at training than the sleep group, *and* there was evidence on the suffix generalisation task that this boosted extraction in the wake group. In other words, in Experiment 5 the extraction in the wake group may have been a consequence of greater initial performance, as opposed to a consequence of wake-dependent *consolidation*. However, the same might not be true for Experiment 4, given the lack of indication in the session two analyses that performance was affected by the initial acquisition level. Consequently, it might be important to address the *possibility* that wake-dependent consolidation, but not sleep-dependent consolidation *might* boost extraction, in the context of previous findings in the literature.

While we have reviewed some studies suggesting that sleep may promote the extraction of regularities (e.g. Gomez et al., 2006), overall this result has not been unanimous in the sleep literature. For instance, some studies have found that consolidation is purely time-dependent, occurring regardless of sleep (e.g. Nemeth et al., 2010; Robertson, Pascual-Leone & Press, 2004). Moreover, there is even evidence to suggest that wake-dependent consolidation specifically, (as opposed to sleep) may benefit consolidation processes (e.g. Song, Howard & Howard, 2007; Werchan & Gomez, 2014).

Given that the wake group's memory for individual, arbitrary mappings was typically lower, this may be another possible explanation for why wake benefited extraction, but sleep did not. For instance, there is evidence both with infants (Werchan & Gomez, 2014) and also with adults (Hennies, Lewis, Durrant, Cousins &ss Lambon Ralph, 2014) that wake facilitates the process of generalization, potentially at the expense of the individual elements. We have been suggesting that sleep actively *protects* individual memories from loss. However, in light of Werchan and Gomez's (2014) findings, there may be another way of looking at our findings. It may be that wake actively facilitates the forgetting of individual items, in order to boost generalization of items. This is because the stabilization of individual details may prevent regularities from being extracted (e.g. Vlach et al., 2012). Systematic mappings (by their nature) appear more frequently in the input in comparison to the less frequently appearing arbitrary mappings. The more frequently appearing, systematic mappings may be tagged for wake-dependent consolidation, whereas the less frequently appearing, arbitrary mappings may be tagged for sleep-dependent consolidation (e.g. Werchan & Gomez, 2014).

This tagging may depend on the neural activation during exposure. Arbitrary mappings may be tagged for sleep-dependent consolidation on the basis of their relatively greater hippocampal-dependence in comparison to systematic mappings. Systematic mappings may be tagged for wake-dependent consolidation on the basis of their relative greater neocortical-dependence (e.g. McClelland et al., 1995). Of course we do not have evidence for this in our own Experiments, so this suggestion is purely conjectural and needs to be investigated in future research using fMRI to track neural activation.

Moreover, it is important to bear in mind that the processes of extraction and generalization likely require novel schemas (on the basis of the shared regularity across the set of items) to be constructed (e.g. Ellis, 2006). However, it may be an oversimplification to assume that schema creation and representation always facilitates encoding and consolidation processes. We have *some* indication that wake *might* benefit extraction of regularities, when the memory for individual items is lower (moreover, regardless of sleep in Experiment 6 we saw increased suffix-gender regularity extraction when individual suffix recall was decreased). Consequently, it might be conjecturally suggested that the creation of schemas may be detrimental to the stabilization of individual items.

There is evidence to support this conjectural suggestion (e.g. Sweegers & Talamini, 2014; Sweegers, Coleman, van Poppel, Cox & Talamini, 2015). Specifically, Sweegers and Talamini (2014) found evidence of regularity generalisation ability immediately. A four hour consolidation delay further boosted this generalisation ability; importantly this was regardless of sleep. Following longer-term consolidation (over several weeks), arbitrary elements were forgotten, whereas memory of regularities was maintained. The findings of their study are relevant to our Experiments, given that they used a hippocampally-dependent task (i.e. a face-location task), and we are similarly predicting our arbitrary mappings to be more hippocampally-dependent. In their study it was suggested that stabilization of exact individual items may prevent regularities from being generalized to new items (e.g. Sweegers & Talamini, 2014).

The initial consolidation-dependent improvement for generalization, followed by longer-term maintenance of generalization in light of individual items' loss found by Sweegers and Talamini (2014) may be informative. It might be that individual, arbitrary mappings need to be stabilized *initially* such that their representations can be used in order to parse the regularities. The extraction of regularities may be a continuous process that requires a longer time period (this would also be consistent with McClelland et al.'s 1995 CLS model). Consequently, the forgetting of individual items within representative networks may increase the salience of the shared regularities, making them more detectable. It is important to bear in mind that in our Experiments we were looking only at the initial stage of language learning. It might be conjecturally suggested that initially, individual items are acquired in a fast process which is facilitated by sleep. With time however, had we included a longer period of consolidation (of several weeks) we might have seen maintenance of the regularity extraction, but forgetting of the individual items (as with Sweegers and Talamini, 2014).

The adaptive role of sleep may be in the initial facilitation of individual item stabilization, such that the individual item representations can be used for extraction. Following this, it is time-dependent consolidation (rather than wake or sleep-dependent specifically) that may facilitate the extraction and generalization processes. If the individual items are not rehearsed or retrieved during that period of time consolidation (as with Sweegers and Talamini, 2014) they may be discarded. Consequently, in the presence of individual arbitrary mappings, sleep may *indirectly* facilitate long-term generalisation and extraction by strengthening the initial representations available for later extraction. Finally, it is also important to bear in mind that we did not have a relative measure between training and recall. Consequently, we do not know whether in comparison to training (had recall been measured at training) sleep may have maintained or even improved recall performance.

Finally, in Chapter 1, on the basis of Walker and Stickgold's (2010)'s model of general sleep-dependent, we predicted that arbitrary mappings would be require potentially just *one* night of sleep to be consolidated, and that SWS might underlie this consolidation. We predicted that the extraction and generalization of regularities contained within systematic mappings may require *several* nights of sleep, with a potential consequence being the forgetting of the individual, arbitrary mappings. While we have support of the first prediction (i.e. that a single night of sleep boosts the consolidation of the arbitrary mappings), we have not tested the second prediction. In other words, as a future direction it would be of interest to test whether the extraction and generalization of regularities may require a longer time period (with several nights of sleep) to be facilitated. For instance, in Experiments 1 and 2 while we had evidence of the extraction of the determiner-gender regularities, we did not have evidence of the suffix-gender regularity extraction. Had a longer time period (spanning several nights of sleep) been provided, the extraction of the suffix-gender regularity *might* have emerged as well.

7.2 Main contributions of this thesis

The findings of experiments reported in this thesis contribute firstly to the understanding of the role of sleep in the consolidation of arbitrary mappings and the extraction of regularities contained within systematic mappings. Secondly, our findings contribute to the understanding of how regularities are extracted regardless of sleep. These contributions based on our findings can be summarised as follows:

Sleep benefits the consolidation of individual arbitrary mappings (individual stems and suffixes), but not the extraction of regularities (determiners and suffixes), regardless of the degree of systematicity, amount of training, regardless of the influence of redundancy of multiples cues to the same gender mapping, and regardless of stem variability. This supports the predictions of the CLS (McClelland et al., 1995). Conjecturally, it may be hypothesised that this is because arbitrary mappings are more hippocampally-dependent than systematic mappings. This needs to be investigated by collecting neural evidence as a future direction.

Determiners and suffixes are functionally different. For determiners, full systematicity appears to be sufficient for extraction. This is in line with the Competition model (e.g. Bates & MacWhinney, 1989) suggesting that the strength of a cue may be determined by its systematicity (and frequency). Cues with higher strength are extracted, as seen with fully systematic determiners. (Moreover, when additional training was provided, there was evidence of determiner-gender regularity extraction immediately even when determiners are not fully systematic.)

In contrast, suffixes were not extracted in the presence of determiners, regardless of their degree of systematicity (and regardless of the amount of training). Even when suffixes were fully systematic and determiners were no longer present in the input (removing determiners as the stronger cue to gender, which may have inhibited suffix-gender regularity extraction), there was still no evidence of suffix-gender reuglarity extraction. This suggests that looking at language acquisition purely on the basis of systematicity (i.e. statistical predictability) is an oversimplification. We found that increasing stem variability boosted suffix-gender regularity extraction, at the expense of individual suffix memory. This also suggests that the forgetting of individual elements boosts the generalisation of regularities. Moreover, it suggests that an increased number of exemplars increases the learner's certainty that the suffix-gender regularity will reliably predict gender in *future* language use, as opposed to just in the trained small set of exemplars. When the suffix-gender regularity is shared across a larger set of exemplars, a stronger representation of the suffix-gender mapping is stored, boosting extraction and generalisation.

Overall, it appears that while sleep inevitably plays *some* role in language consolidation processes, its role is ultimately limited. Much more crucial are factors relating to the structure of the language input. Namely, while greater systematicity boosts category learning, it is affected by factors such as firstly the temporal order of information. Information at the beginning of a language item is encountered earlier, and hence may receive more processing for extraction. Secondly, cues which have stronger predictive value may inhibit the extraction of weaker cues. Although this is true to some extent, even making suffixes nonredundant was not sufficient for extraction. Suffixes were only extracted when stem (and hence exemplar) variability was increased. Consequently, the language factors beyond sleep which affect language acquisition include systematicity of the mapping, temporal order of information, redundancy of cues and exemplar variability.

7.3. Future directions

A key prediction across our Experiments has been that sleep should preferentially benefit the consolidation of the more hippocampally-dependent mappings, explaining the benefit of sleep for arbitrary mappings over systematic mappings. This prediction was based on the suggestions of the CLS (e.g. McClelland et al., 1995) and previous findings suggesting that sleep benefits the consolidation of hippocampally-dependent information (e.g. Rasch et al., 2007). However, this was not tested in *our* Experiments. Evidence using fMRI is needed to confirm that as we are predicting, our individual arbitrary mappings are indeed more hippocampally-dependent than our systematic mappings, and whether the degree of hippocampal-dependence can be related to the sleep-dependent benefit.

Secondly, it is of interest to explore the case of intermittent systematicity of the mapping. As we have already discussed in this Chapter, learning occurs not only on trials where a cue is present, but also on trials where a cue is not present. For example, in Experiments 1 and 2, although the same gender could be predicted from either the suffix *"eem"* or the suffix *"esh",* each suffix predicted its corresponding gender 100% of the time. In other words, the suffix *"eem"* was *always* seen paired with the same gender. It would be of interest to investigate learning where each suffix maps on to its corresponding gender intermittently (as opposed to 100% of the time). One way of doing this is by not having the suffix-gender regularity present on every trial. For example, if the regular suffixes in the input include *"eem"* and *"esh"* mapping on to pictures depicting characters of a female gender, then on some trials depicting female gender pictures neither suffix would be present. On the basis of the Competition model (e.g. Bates & MacWhinney, 1989), we would expect that mappings are less likely to be extracted in the case of intermittent systematicity. Given that in our Experiments we have found that the degree of systematicity influenced determiner and suffix mappings differently (i.e. full systematicity of the mapping was sufficient for determiners to be extracted, but not for suffixes), it is of interest whether intermittent systematicity would also influence determiner and suffix mappings differently. This would further support our suggestion that determiners and suffixes might be processed for extraction differently.

Appendices

Appendix 1: control measures for Experiment 1

Method

The following description of the control measures applies to all of the six Experiments presented in this thesis.

Controlling for group differences: Questionnaires and vigilance measures

The Stanford Sleepiness questionnaire (SSS; Hoddes, Zarcone, Smythe, Philips, & Dement, 1973) was given to participants to complete at the start of both sessions to measure levels of sleepiness. Specifically, participants rated their sleepiness on a scale of 1-7. Participants also rated how they felt at the current moment on a scale of 1 (feeling nearly absent) to 7 (most intense feeling) in terms tiredness, drowsiness, motivation, and concentration by ticking the applicable box. The sleep habits questionnaire and the Epworth test were used to screen for sleep disorders.

Vigilance task

This task was used to measure participants RTs as indication of their alertness (i.e.) vigilance level at the start of each session. There were 18 trials. First, a fixation cross appeared on the screen for 500 ms, then a ‘1’ and ‘0’ appeared simultaneously on opposite sides of the screen until the participant responded. On 9 out of 18 trials the ‘1’ appeared on the left hand-side of the screen (as the ‘0’ appeared on the right) and vice versa for the remaining 9 trials (‘0’ on the left and ‘1’ on the right). The presentation order (whether the ‘1’ was on the right or on the left) was randomized. Participants were asked to use the buttonbox to respond if the ‘1’ was on the right or on the left.

Results

These control (vigilance) measures were compared between the two sessions and the two groups. Six mixed ANOVAs were run on the dependent variables of sleepiness (i.e. SSS), tiredness, drowsiness, motivation, concentration and vigilance RTs, each time with the between subjects variable of group (sleep or wake) and the within subjects variable of session (session one or two). There were no significant differences between groups or between sessions, nor a group x session interaction, in the self-reported levels of sleepiness (i.e. SSS), tiredness or drowsiness (see Table 19 for ANOVA results). There was a marginally significant main effect of group for the measure of concentration, with the wake group having overall higher levels of concentration than the sleep group across both sessions (wake group mean = 5.11; sleep group mean = 4.67). There was a significant main effect of group for motivation, with the wake group being more motivated than the sleep group across both sessions (wake group mean = 4.78; (sleep group mean = 4.22).

Table 19. Results of mixed ANOVAs for control measures in Experiment 1.

	F	df- value	*p* - value
		SSS	
Group	2.25	1, 52	.139
Session	.21	1, 52	.653
Group x Session	1.39	1, 52	.244
		Tiredness	
Group	.02	1, 52	.888
Session	.02	1, 52	.878
Group x Session	.02	1, 52	.878
		Drowsiness	
Group	.19	1, 52	.662
Session	.00	1, 52	1.000
Group x Session	.03	1, 52	.864
		Motivation	
Group	4.48	1, 52	.039
Session	2.55	1, 52	.116
Group x Session	1.44	1, 52	.236
		Concentration	
Group	3.59	1, 52	.064
Session	1.18	1, 52	.283
Group x Session	3.27	1, 52	.077
		Vigilance task RTs	
Group	3.35	1, 51	.073
Session	4.67	1, 51	.035
Group x Session	5.29	1, 51	.026

*NB. One participant in the wake group was excluded from the vigilance task analysis due to not following instructions

As shown in Figure 76, there were no overall differences in reaction times between participants in the two groups on the vigilance task, but the group x session interaction suggests that the difference in vigilance between sessions is driven by a larger change in the wake group reaching vigilance levels of the sleep group in Session 2 (see Table 19 for ANOVA results).

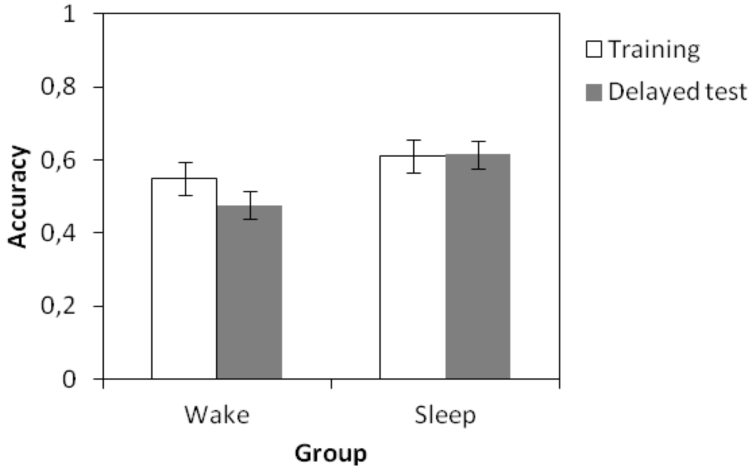


Figure 76. Vigilance task performance of both groups for both sessions in Experiment 1.

Two between subjects t-tests were run to measure group differences on firstly the Epworth questionnaire total (measuring typical daytime sleepiness and likelihood of sleep disorders) and secondly on typical sleep length. There were no differences between participants in the sleep and wake groups on these measures (see Table 20 for results).

Table 20. Results of between subjects t-tests of control measures in Experiment 1.

	t	df - value	*p* – value
Epworth Total	1.14	52	.262
Sleep length	.52	52	.602

These findings suggest that any sleep benefits found on the language tasks in Experiment 1 cannot be as a result of circadian differences or as a result of differences in motivation, concentration, tiredness, drowsiness, sleepiness or vigilance between participants in the two groups (especially as participants in the wake group were more motivated leading to expect that participants in the wake group would have performed better). Typical daytime sleepiness (measured by the Epworth questionnaire total) and typical sleep length were also similar for participants in both groups and therefore also should not affect group differences on the main tasks.

Appendix 2: control measures for Experiment 2

Six independent samples t-tests with the independent variable of group (Experiment 1 sleep/ Experiment 2 sleep) were ran to compare session two alertness levels. This was done in order to comment on the similarities between the two sleep in the lab versus sleep at home groups (see Table 21 for results of independent-subjects t-tests).

In sleepiness (as measured by Stanford Sleepiness Scale), participants in Experiment 1 (mean = 2.85) were higher on sleepiness than participants in Experiment 2 (mean = 2.00). In tiredness, participants in Experiment 1 (mean = 3.26) were marginally more tired than participants in Experiment 2 (mean = 2.47). In motivation, participants in Experiment 1 (mean = 3.96) had lower motivation than participants in Experiment 2 (mean = 5.76). In concentration, participants in Experiment 1 (mean = 4.37) had lower concentration than participants in Experiment 2 (mean = 5.71). In addition, regardless of session two, on the Epworth test measuring typical daytime sleepiness, participants in Experiment 1 (mean = 5.41) had a lower score than participants in Experiment 2 (mean = 7.35), indicating a lower typical level of sleepiness. Participants from both Experiments were in the normal range on the Epworth test.

Overall, at session two, participants in Experiment1 (in comparison to participants from Experiment 2) had higher sleepiness (as measured by SSS), lower concentration, more tiredness and less motivation. This suggests that potentially sleeping in the lab may have boosted participants' alertness levels. The more controlled sleeping conditions in Experiment 2 may have made participants feel more rested.

Table 21. Results of independent samples t-tests comparing session two alertness between Experiment 1 and 2.

	t	df-value	*p* - value
SSS	2.44	42	.019
Tiredness	1.81	42	.078
Drowsiness	1.05	42	.301
Motivation	-5.44	42	<.001
Concentration	-3.56	42	<.001
Vigilance task (RTs)	2.75	42	.014
Epworth	-2.13	42	.039
Sleep length	-.65	42	.522

Appendix 3: control measures for Experiment 3

As with Experiment 1, six mixed ANOVAs were run on the dependent variables of sleepiness (i.e. SSS), tiredness, drowsiness, motivation, concentration and vigilance RTs, each time with the between subjects variable of group (sleep or wake) and the within subjects variable of session (session one or two). There were no significant differences between groups or between sessions, nor a group x session interaction, in the self-reported levels of drowsiness, motivation or concentration (see Table 22 for results of the ANOVAs).

There was a significant main effect of group for sleepiness (as measured by SSS), with participants from the wake group reporting lower levels of sleepiness than participants in the sleep group (wake group overall mean = 2.20, sleep group overall mean = 2.64). There was a significant main effect of group for tiredness, with participants from the wake group reporting lower levels of tiredness than participants in the sleep group (wake group overall mean = 2.58, sleep group overall mean = 3.20; Table 22). Overall, according to self-reported questionnaires, participants from the wake group may have been more alert than participants from the sleep group (based on the lower self-reported levels of sleepiness and tiredness).

Table 22. Results of mixed ANOVAs for control measures in Experiment 3.

	F	df- value	*p* - value
		SSS	
Group	4.26	1,48	.044
Session	.06	1,48	.807
Group x Session	1.50	1,48	.226
		Tiredness	
Group	5.26	1,48	.026
Session	.40	1,48	.533
Group x Session	3.55	1,48	.066
		Drowsiness	
Group	1.71	1,48	.197
Session	.22	1,48	.644
Group x Session	1.46	1,48	.233
		Motivation	
Group	2.31	1,48	.135
Session	.86	1,48	.357
Group x Session	.01	1,48	.933
		Concentration	
Group	3.01	1,48	.089
Session	.64	1,48	.429
Group x Session	.16	1,48	.692
		Vigilance task RTs	
Group	.47	1,47	.499
Session	21.66	1,47	<.001
Group x Session	1.01	1,47	.321

*NB. One participant in the wake group was excluded from the vigilance task analysis due to software malfunction.

On the vigilance RT task, there was a significant main effect of session (see Table 22 for results of the ANOVA). As shown in Figure 77, all participants were more vigilant in session 2 (than in session 1), regardless of group. This is because all participants had slower RTs in session 1 than in session 2.

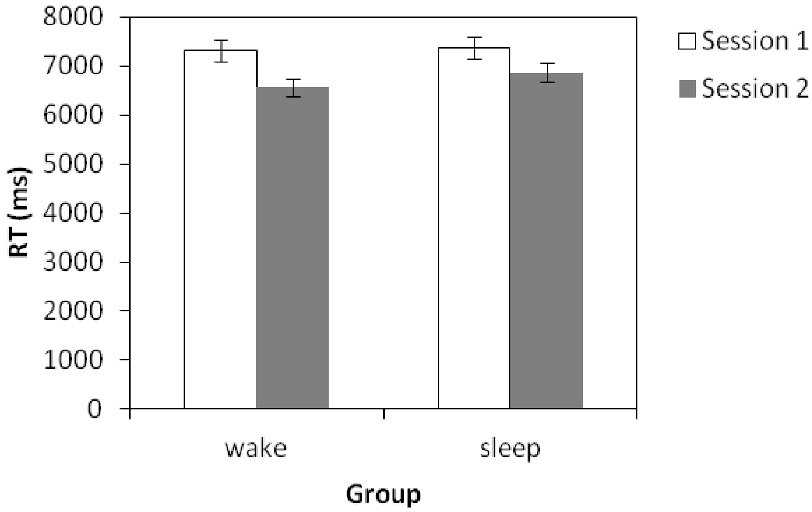


Figure 77. Vigilance task performance of both groups for both sessions in Experiment 3.

As with Experiment 1, two between subjects t-tests were run to measure differences between the two groups on firstly the Epworth questionnaire (measuring typical daytime sleepiness and screening for sleep disorders) and secondly on typical sleep length. There were no differences between participants in the sleep and wake groups on these measures (see Table 23 for results).

Table 23. Results of between subjects t-tests of control measures in Experiment 3.

	t	df – value	*p* – value
Epworth Total	-.47	48	.638
Sleep length	-1.11	48	.274

Overall, these findings suggest that the sleep benefits found in the suffix recall task in Experiment 3 cannot be as a result of circadian differences or as a result of differences in motivation, concentration, tiredness, drowsiness, sleepiness or vigilance between participants in the two groups (especially as participants in the wake group felt less tired and expressed feeling lower levels of sleepiness compared to participants in the sleep group, leading to expect that participants in the wake group would have performed better). Typical daytime sleepiness (measured by the Epworth questionnaire total) and typical sleep length were also similar for participants in both groups and therefore also should not affect group differences on the main tasks.

Appendix 4: control measures for Experiment 4

In Experiment 4 session two language tasks we analysed the data for the delayed groups separately from the data of the immediate group. Consequently, the analysis of the control measures in Experiment will follow the same structure. Firstly, the control measures of the two delayed groups will be compared between the two sessions. Secondly, the two delayed groups' and the immediate group's control measures at session two will be compared qualitatively, rather than statistically. This is because the immediate group was tested using session two tasks *immediately* after session one tasks, whereas the two delayed groups had a 12-hour delayed in between the two sessions. Hence, this should result in higher fatigue levels (due to a longer overall session) in the immediate group. The typical sleep length and the Epworth score however *will* be compared across the three groups statistically, given that these two general measures are not dependent on the specific vigilance levels at the time of the session.

To analyse the control measures for the two delayed groups, as with Experiments 1 and 3, six mixed ANOVAs were run on the dependent variables of sleepiness (i.e. SSS), tiredness, drowsiness, motivation, concentration and vigilance RTs, each time with the between subjects variable of group (sleep or wake) and the within subjects variable of session (session one or two). There were no significant differences between groups or between sessions, nor any group x session interactions, in the self-reported levels of sleepiness (as measured by SSS), tiredness, drowsiness, or concentration (see Table 24 for results of the ANOVAs). For motivation, all participants had higher self-reported levels of motivation in session 1 (session 1 mean = 5.14) than in session 2 (session 2 mean = 4.62), regardless of group.

Table 24. Results of mixed ANOVAs for control measures in Experiment 4.

	F	df- value	*p* - value
		SSS	
Group	1.99	1,48	.165
Session	.86	1,48	.360
Group x Session	.26	1,48	.610
		Tiredness	
Group	1.97	1,48	.167
Session	1.08	1,48	.304
Group x Session	1.08	1,48	.304
		Drowsiness	
Group	.36	1,48	.553
Session	.07	1,48	.794
Group x Session	.08	1,48	.930
		Motivation	
Group	.07	1,48	.791
Session	7.23	1,48	.010
Group x Session	.39	1,48	.538
		Concentration	
Group	.02	1,48	.878
Session	2.18	1,48	.146
Group x Session	.03	1,48	.854
		Vigilance task RTs	
Group	1.53	1,48	.223
Session	20.23	1,48	<.001
Group x Session	1.37	1,48	.247

On the vigilance RT task, all participants showed higher vigilance in session 2, than in session 1, regardless of group, as shown in Figure 78. This is because participants had slower RTs in session 1 than in session 2.

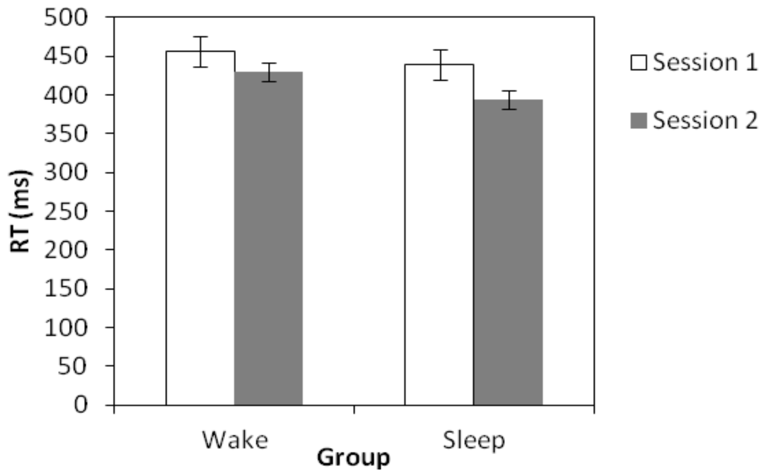


Figure 78. Vigilance task performance of both of the delayed groups for both sessions in Experiment 4.

Overall these results suggest that any indications of differences found on the session two language tasks between the delayed groups in Experiment 4 cannot be as a result of differences in vigilance levels between the two delayed groups.

Next, the typical sleep length and Epworth score total were compared for the two delayed groups *and* the immediate group, to provide indication of the general normality of sleep in the three groups. A one way between subjects ANOVA was run with the independent variable of group (sleep, wake or immediate). The dependent variable was firstly Epworth total, and secondly typical sleep length. As shown in Table 25, there were no differences in Epworth total or typical sleep length across the three groups of participants (i.e. from the wake, sleep and the immediate groups). Hence, there was no indication that differences in the typical normality of sleep could have affected performance of the three groups (e.g. on the training task) in Experiment 4.

Table 25. Results of the one-way ANOVAs (for the three groups) of control measures in Experiment 4.

	F	df – value	*p* – value
Epworth Total	.03	1,98	.971
Sleep length	.55	1,98	.581

Numerically, as shown in Table 26, there was most importantly indication that the immediate group had higher levels of sleepiness (as measured by the SSS) than the two delayed groups. Numerically, the immediate group also had higher levels of drowsiness than the two delayed groups. Overall, numerically there was indication that the immediate group was overall more fatigued than the two delayed groups at the session two tests. The immediate group being more fatigued at session two tests was the reason why the immediate group's session two language tasks were analysed separately from the two delayed group's.

Table 26. Means for the session two control measures for the three groups in Experiment 4. Parentheses denote the standard deviation.

	Wake group	Sleep group	Immediate group
SSS	2.52 (1.19)	2.68 (1.07)	3.18 (1.32)
Tiredness	3.68 (1.28)	3.08 (1.52)	3.67 (1.51)
Drowsiness	2.44 (1.19)	2.60 (1.29)	3.25 (1.57)
Motivation	4.52 (1.50)	4.72 (1.34)	4.41 (1.17)
Concentration	4.76 (1.20)	4.76 (1.51)	4.41 (1.15)

Session 2 Stanford Sleepiness Scale (SSS) comparison

It was crucial to see whether participants in the immediate groups could have been more fatigued than participants in the delayed groups (as a result of the immediate groups having had less rest break before the session 2 tests). In order to investigate this, a two way between subjects ANOVA was run with the independent variables of consolidation type (immediate or delayed) and time of day (AM or PM, note that the delayed AM-tested group is also the sleep group, as session two took place in the morning for the sleep group). The dependent variable was the average SSS, i.e. the participant's sleepiness rating at session two.

There was a significant main effect of consolidation type ($F(1,100) = 6.73$, $p = .011$). There were a marginally significant main effect of time of day ($F(1,100) = 2.85$, $p = .095$). There was a significant interaction between consolidation type and time of day ($F(1,100) = 8.57$, $p = .004$). As shown in Figure 79, participants tested immediately had higher self-ratings of sleepiness than participants tested after a delay (i.e. the wake and sleep groups), as evidenced by the main effect of consolidation type. As evidenced by the *marginally* significant interaction between consolidation type and time of day, participants from the PM-immediate group had *marginally* higher ratings of sleepiness than participants from the AM-immediate group. It was the other way around in the two delayed groups: participants from the PM-delayed group (i.e. the wake group) had *marginally* lower ratings of sleepiness than participants from the AM-delayed group (i.e. the sleep group). Hence, there was no consistent pattern to suggest a specific time of day effect. This analysis provided evidence to suggest that participants tested immediately were more fatigued than participants tested following a delay. Given that sleepiness differed between participants tested immediately compared to participants tested following a delay (whereby the participants tested immediately were more fatigued), data from the two sets of participants was analysed separately.

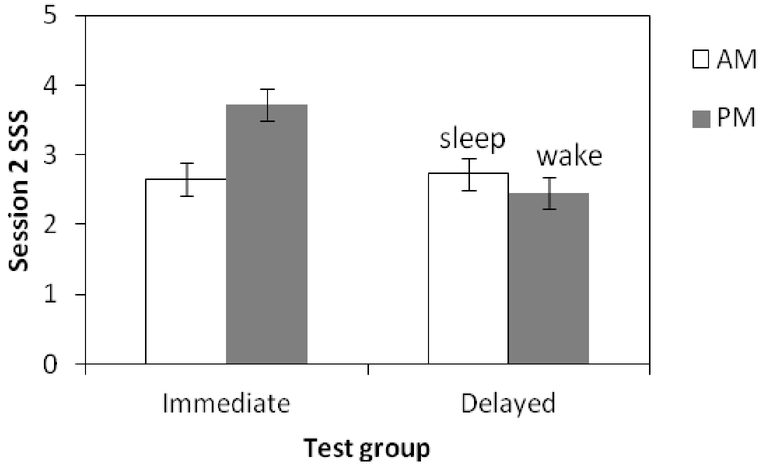


Figure 79: Session 2 SSS (sleepiness) differences

Session two language tasks: the two immediate groups

An analysis comparing the performance of the two immediate control groups on the session two language tasks is included here, to supplement the analysis presented in Chapter 4. This is done in order to address questions of time-of-day effects, i.e. whether the consolidation of individual arbitrary mappings, or the extraction of the regularities contained within systematic mappings might have been affected by time-of-day.

Most arbitrary mapping: vocabulary

Translation recognition task

As already described in Chapter 4, participants' memory for the trained words and their meanings was tested using a translation recognition task. Half of the trials contained the match (correct) English word-novel word pairing translations, half contained mismatch (incorrect) pairings. As already explained in Chapter 4, to account for the potential differences in participants' ability to discriminate between correct and incorrect pairings, the non-parametric (A') discrimination formula (Donaldson, 1992) was used to analyse the data from the delayed groups.

A between-subjects t-test was run with the independent variable of group (AM-immediate, or PM-immediate), and A' discrimination as the dependent variable. There were no significant differences between the two groups ($t(49) = .49$, $p = .626$). Participants from both of the immediate groups had similarly good performance on the translation recognition task, as shown in Figure 80. Hence, this analysis provided no evidence for time of day effects.

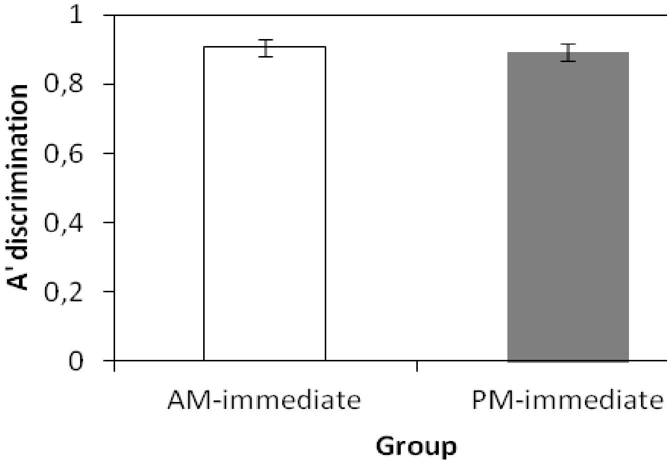


Figure 80. Performance of the immediate groups on the translation recognition task.

Recall (stems)

As already explained in Chapter 4, participants' memory for the trained word-picture mappings was tested using a picture naming recall task. In line with the analysis presented in Chapter 4, data from this task was analysed for stems and suffixes separately. To compare stem recall between the two immediate groups, a two-way mixed ANOVA was run. The between-subjects variable was group (AM-immediate/PM-immediate), the within-subjects variable was cue type. Stem recall accuracy (proportion correct out of total productions) was the dependent variable.

There was no significant main effect of group ($F(1, 49) = 2.87$, $p = .097$). There was a significant main effect of cue type ($F(1, 49) = 4.48$, $p = .039$). There was no significant interaction between group and cue type ($F(1, 49) = .23$, $p = .632$). As shown in Figure 81, the stem recall of all participants benefited from the grapheme cue more than the picture only being present. Although numerically the performance of participants in the AM-immediate group was overall higher than the performance of participants in the PM-immediate group, this was not significant.

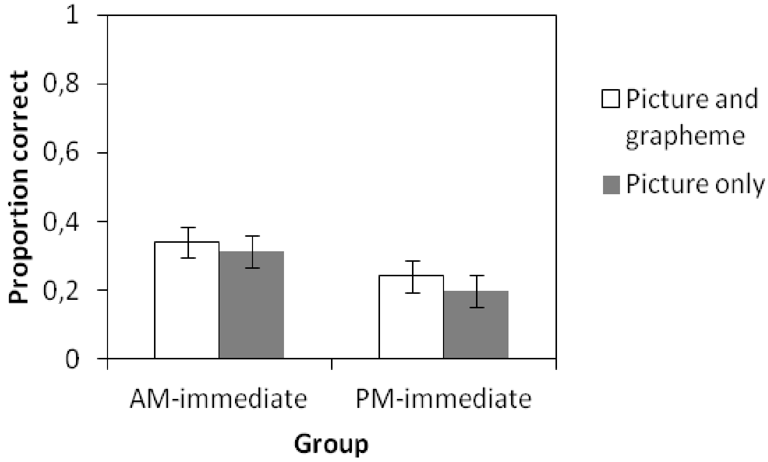


Figure 81. Stem recall accuracy for the immediate groups.

Overall, the results from the vocabulary tests did not provide any significant evidence for time of day effects.

Most systematic mapping: suffixes (grammar)

For the suffix tests, it was important to test whether any differences as a result of time of day emerged in the consolidation of the suffix-picture mappings and in the extraction of suffix-gender regularities. This was analysed by comparing the performance of the two immediate groups on both the suffix recall task, and on the suffix generalisation task. As with the analysis of the delayed groups in Chapter 4, we were particularly interested in looking at the main effect of consistency on the suffix generalisation task as indication of suffix-gender regularity extraction.

Recall (suffixes)

As with Chapter 4, to assess participants' memory of the systematic element of the word, the recall accuracy of the suffixes (which were separated from stems in the transcriptions) was analysed.

For the immediate groups, a two-way mixed ANOVA was run. The between subjects variable was group (AM-immediate/PM-immediate), the within-subjects variable was cue type (picture and grapheme, or picture only). Suffix recall accuracy (proportion correct out of total productions) was the dependent variable. There was a marginally significant main effect of group ($F(1, 49) = 3.32$, $p = .074$). There was no significant main effect of cue type ($F(1, 49) = .09$, $p = .760$). There was no significant interaction between group and cue type ($F(1, 49) = .61$, $p = .438$).

As shown in Figure 82, participants had similar performance regardless of the cue type. In addition, participants in the AM-immediate group had marginally higher performance for suffix recall accuracy than participants in the PM-immediate group. However, there was no *strong* statistical evidence that time of day affected the consolidation of the trained suffix-picture mappings.

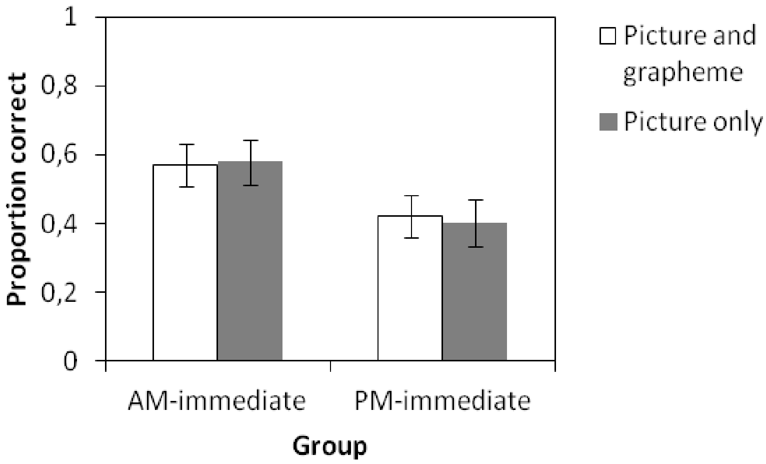


Figure 82. Suffix accuracy on the recall task for the immediate groups.

Suffix generalisation task

As already explained in Chapter 4, to test if participants extracted the suffix-gender regularities which they were exposed to during training, a suffix generalisation task was used. Crucially, the pattern of endorsement of consistent versus inconsistent trials was of interest here, as opposed to the overall level of performance.

For the immediate groups, a two-way mixed ANOVA was run. The between subjects variable was group (AM-immediate/PM-immediate), the within-subjects variable was consistency (items consistent versus inconsistent with the trained regularities). Proportion endorsed was the dependent variable. There was no significant main effect of group ($F(1, 49) = 1.99$, $p = .165$) and no significant main effect of consistency ($F(1, 49) = .23$, $p = .636$). There was no significant interaction between group and consistency ($F(1, 49) = 1.45$, $p = .234$). As shown in Figure 83, participants had a similar level of endorsement regardless of consistency; hence there was no evidence of suffix-gender regularity extraction, regardless of group. Consequently, this task provided no evidence that time of day affected suffix-gender regularity extraction.

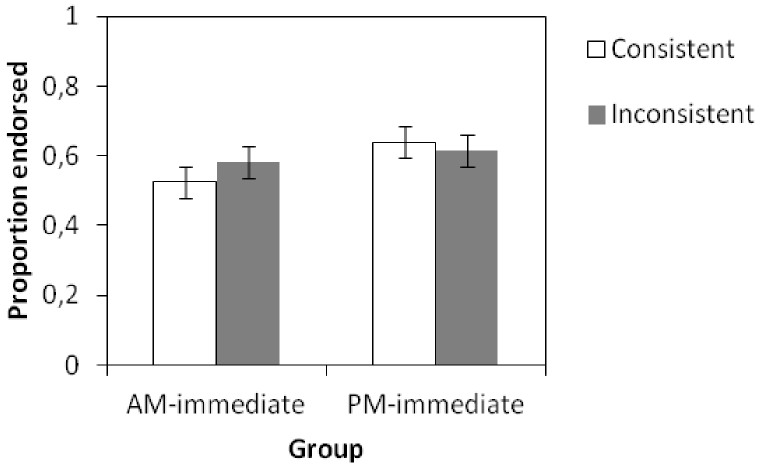


Figure 83. Suffix generalisation performance for the immediate groups.

As with Chapter 4, one sample t-tests against chance were also run, separately for consistent and inconsistent items.

The results of one-sample t-tests against chance (0.5) are shown in Table 27. As a reminder, the pattern that would suggest that the suffix-gender regularities were extracted, would be if consistent items were endorsed significantly above chance, and inconsistent items were endorsed significantly below chance (or at least *at* chance). This pattern was not shown in either of the groups, hence there was no evidence that the suffix-gender regularities were extracted. More specifically, the AM-immediate group endorsed consistent items at chance, and inconsistent items marginally above chance. The PM-immediate group endorsed both consistent and inconsistent items significantly above chance.

Table 27. Results of one sample t-tests against chance on the suffix generalisation task.

	Word-picture gender pairing	t	df value	*p* - value
AM-immediate	Consistent	.68	25	.506
	Inconsistent	2.02	25	.055
PM-immediate	Consistent	2.97	24	.007
	Inconsistent	2.45	24	.022

To summarize, the findings from the suffix tests suggest that participants from the immediate groups reached a similar level of suffix recall accuracy, hence there was no evidence that time of day affected suffix consolidation. In addition, there was no evidence of suffix-gender regularity extraction immediately, and there was no evidence that time of day affected suffix-gender regularity extraction.

Intermediate systematicity mapping: determiners (grammar)

Determiner selection task

As already outlined in Chapter 4, participants' memory for the trained determiner-picture mappings was assessed using a determiner selection task. For the immediate groups, a between subjects t-test was run with the independent variable of group (AM-immediate or PM-immediate). Proportion correct was the dependent variable. There were no significant differences between the two groups (t(49)= -.85, p = .400).

As shown in Figure 84, both of the immediate groups reached a similarly low level of accuracy. There was no evidence that time of day affected the memory for trained determiner-picture mappings.

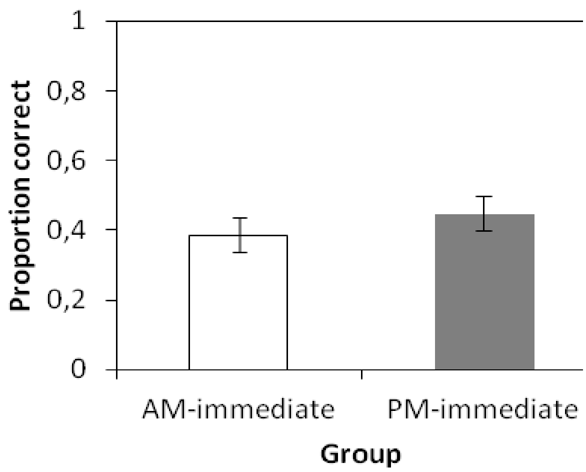


Figure 84. Performance on the determiner selection task.

The results of one-sample t-tests against chance (set at 0.25, given that there were 4 determiners), are shown in Table 28. The results indicated that participants from both of the immediate groups were above chance at correctly selecting determiner-picture mappings. This was also in line with the against chance results of the two delayed groups presented in Chapter 4.

Table 28. Results of Experiment 4 one sample t-tests against chance on the determiner selection task.

	t	df value	*p* - value
AM-immediate	2.51	25	.019
PM-immediate	4.29	24	<.001

Determiner and suffix generalisation task

As already outlined in Chapter 4, as an additional test of the extraction of the determiner-gender regularities, the determiner and suffix generalisation task was used. As a reminder, in this task we were interested in the *pattern* of endorsement, as opposed to overall performance.

For the immediate groups, a two-way mixed ANOVA was run. The between subjects variable was group (AM-immediate/PM-immediate), the within-subjects variable was consistency (items consistent versus inconsistent with the trained regularities). Proportion endorsed was the dependent variable. There was a significant main effect of group ($F(1, 49) = 4.14$, $p = .047$). There was a significant main effect of consistency ($F(1, 49) = 5.94$, $p = .019$). There was no significant interaction between group and consistency ($F(1, 49) = .94$, $p = .337$). As shown in Figure 85, participants from both groups showed extraction of the trained determiner-gender regularities, as evidenced by the main effect of consistency. Crucially, as there was no significant group x consistency interaction, the two groups had a similar *pattern* of endorsement. As the extraction ability is indicated by the *pattern* of endorsement, as opposed to the overall performance, there was no evidence that time of day affected the extraction of determiner-gender regularities.

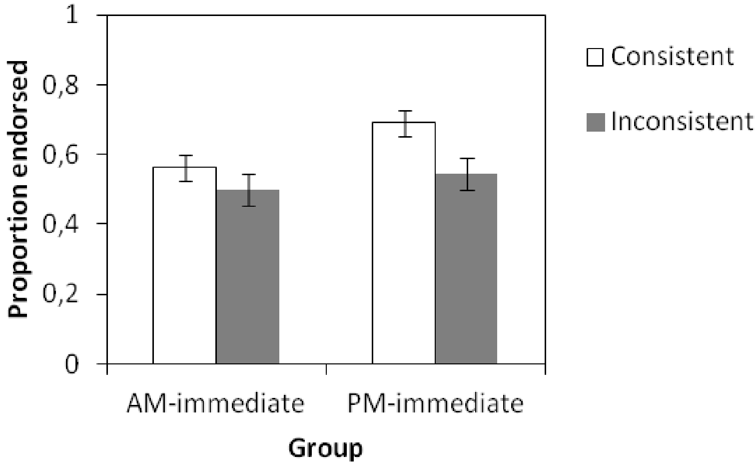


Figure 85. Determiner and suffix generalisation task performance for the immediate groups.

As with Chapter 4, performance was additionally compared against the chance level of 0.5. As a reminder, if determiner-gender regularities are extracted, the expected pattern is that participants would be significantly above chance on the consistent items, and either at chance or significantly below chance on the inconsistent items. As shown in Table 29, participants in the AM-immediate group were at chance for endorsing both the consistent and the inconsistent items. Participants in the PM-immediate group were above chance for endorsing consistent items, and at chance for endorsing inconsistent items. Consequently, based on one sample t-tests against chance, the performance of the AM-immediate group was *not* in line with determiner-gender regularity extraction, while the performance of the PM-immediate group *was* in line with determiner-gender regularity extraction. However, this finding needs to be interpreted with caution, given that there was no significant group x consistency interaction in the ANOVA, and hence no strong evidence of differences between groups on this task.

Table 29. Results of one-sample t-tests against chance on the determiner and suffix generalisation task.

	Word-picture gender pairing	t	df value	*p* - value
AM-immediate	Consistent	1.63	25	.116
	Inconsistent	.04	25	.967
PM-immediate	Consistent	5.68	24	<.001
	Inconsistent	.95	24	.352

To summarise the findings from the determiner tests, the determiner selection task analysis provided evidence that memory for individual determiner-picture mappings was low. There was no evidence that time of day affected the consolidation of trained determiner-picture mappings. The determiner and suffix generalisation task analysis provided evidence that all participants extracted the determiner-gender regularities, regardless of group. There was no strong evidence that time of day affected determiner-gender regularity extraction (given that there was no significant group x consistency interaction).

Appendix 5: control measures for Experiment 5

As with the previous Experiments, six mixed ANOVAs were run on the dependent variables of sleepiness (i.e. SSS), tiredness, drowsiness, motivation, concentration and vigilance RTs, each time with the between subjects variable of group (sleep or wake) and the within subjects variable of session (session one or two).

Participants from the sleep group had significantly higher self-reported levels of sleepiness (as measured by the SSS, than participants from the wake group (average sleepiness: sleep group mean = 2.84, wake group mean = 2.23). Participants from the wake group had significantly higher self-reported levels of motivation than participants from the sleep group (wake group mean = 5.31, sleep group mean = 4.42). Participants from the wake group also had higher self-reported levels of concentration than participants from the sleep group (wake group mean = 5.29, sleep group mean = 4.42), the results of the ANOVAs are shown in Table 30. There were no other significant differences between groups on the *self-reported* control vigilance measures.

Table 30. Results of mixed ANOVAs for control measures in Experiment 5.

	F	df- value	*p* - value
		SSS	
Group	8.47	1,47	.006
Session	.47	1,47	.495
Group x Session	.80	1,47	.376
		Tiredness	
Group	.29	1,47	.593
Session	.54	1,47	.467
Group x Session	.01	1,47	.905
		Drowsiness	
Group	.21	1,47	.647
Session	2.73	1,47	.105
Group x Session	2.73	1,47	.105
		Motivation	
Group	10.6	1,47	.002
Session	.17	1,47	.680
Group x Session	.04	1,47	.842
		Concentration	
Group	11.27	1,47	.002
Session	.20	1,47	.656
Group x Session	.69	1,47	.412
		Vigilance task RTs	
Group	.01	1,47	.923
Session	16.30	1,47	<.001
Group x Session	.11	1,47	.747

For the vigilance RT task, as shown in Figure 86, all participants had faster reaction times at session 2, suggesting that regardless of group, all participants were more vigilant at session 2 than at session 1.

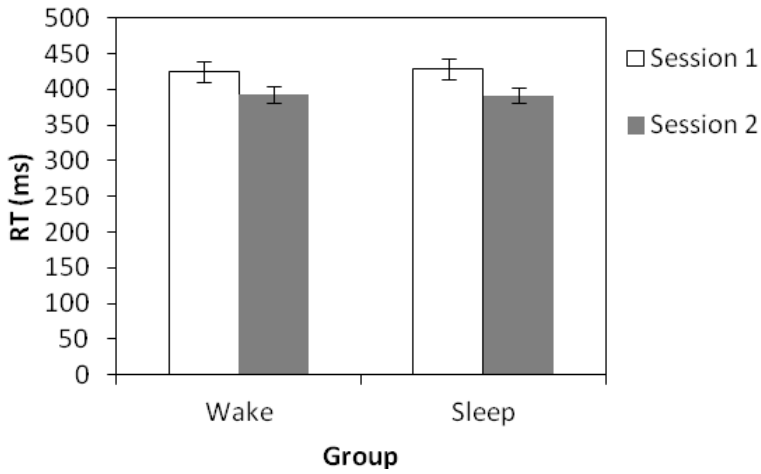


Figure 86. Vigilance task performance of both groups for both sessions in Experiment 5.

As with the previous Experiments, two between subjects t-tests were run to measure differences between the two groups on firstly the Epworth questionnaire (measuring typical daytime sleepiness and screening for sleep disorders) and secondly on typical sleep length. Participants from the sleep group had significantly higher self-reported levels average sleep length (see Table 31 for results of the t-test) than participants from the wake group (average sleep length: sleep group mean = 8.36, wake group mean = 7.77 hours per night). There were no differences on the Epworth total.

Table 31. Results of between subjects t-tests of control measures in Experiment 5.

	t	df – value	*p* – value
Epworth Total	-.70	47	.485
Sleep length	-2.68	47	.010

Overall, participants from the sleep group had higher levels of sleepiness and lower levels of motivation and concentration than participants from the wake group in Experiment 5.

Appendix 6: control measures for Experiment 6

As with the previous Experiments, six mixed ANOVAs were run on the dependent variables of sleepiness (i.e. SSS), tiredness, drowsiness, motivation, concentration and vigilance RTs, each time with the between subjects variable of group (sleep or wake) and the within subjects variable of session (session one or two).

All of the participants had significantly higher self-reported levels of motivation in session 2 (mean = 4.84) than in session 1 (4.38), regardless of group. There were no other significant differences between groups or between sessions on the *self-reported* control vigilance measures, as shown in Table 32. This suggests that participants from both groups had comparable levels of vigilance, and therefore vigilance levels were unlikely to influence performance on the language tasks.

Table 32. Results of mixed ANOVAs for control measures in Experiment 6.

	F	df- value	*p* - value
		SSS	
Group	1.30	1,47	.261
Session	1.75	1,47	.193
Group x Session	1.19	1,47	.282
		Tiredness	
Group	.08	1,47	.783
Session	.04	1,47	.839
Group x Session	.34	1,47	.563
		Drowsiness	
Group	2.48	1,47	.122
Session	.09	1,47	.764
Group x Session	.09	1,47	.764
		Motivation	
Group	.01	1,47	.946
Session	6.42	1,47	.015
Group x Session	.17	1,47	.683
		Concentration	
Group	.03	1,47	.856
Session	1.27	1,47	.265
Group x Session	2.43	1,47	.126
		Vigilance task RTs	
Group	.48	1,47	.493
Session	34.14	1,47	<.001
Group x Session	.24	1,47	.625

Participants from both groups also had similar RTs on the vigilance task. All participants were faster at Session 1 than at Session 2 (which is the opposite to the session differences found in Experiment 5), regardless of group, as shown in Figure 87.

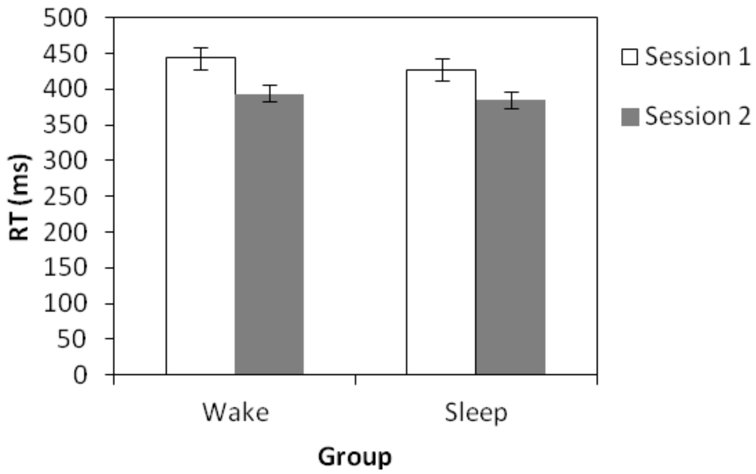


Figure 87. Vigilance task performance of both groups for both sessions in Experiment 6.

As with the previous Experiments, two between subjects t-tests were run to measure differences between the two groups on firstly the Epworth questionnaire (measuring typical daytime sleepiness and screening for sleep disorders) and secondly on typical sleep length. As shown in Table 33, there were no significant differences on the Epworth total or typical sleep length between the wake and the sleep groups.

Table 33. Results of between subjects t-tests of control measures in Experiment 6.

	t	df – value	*p* – value
Epworth Total	.61	47	.542
Sleep length	-0.26	47	.800

Overall, these findings suggest that any sleep benefits found on the recall task in Experiment 6 cannot be as a result of circadian differences or as a result of differences in motivation, concentration, tiredness, drowsiness, sleepiness or vigilance between participants in the two groups.

References

Albouy, G., Fogel, S., King, B. R., Laventure, S., Benali, H., Karni, A., ... & Doyon, J. (2015). Maintaining vs. enhancing motor sequence memories: respective roles of striatal and hippocampal systems. *Neuroimage*, *108*, 423-434.

Albouy, G., King, B. R., Maquet, P., & Doyon, J. (2013). Hippocampus and striatum: Dynamics and interaction during acquisition and sleep‐related motor sequence memory consolidation. *Hippocampus*, *23*(11), 985-1004.

Albouy, G., Sterpenich, V., Balteau, E., Vandewalle, G., Desseilles, M., Dang-Vu, T., ... & Peigneux, P. (2008). Both the hippocampus and striatum are involved in consolidation of motor sequence memory. *Neuron*, *58*(2), 261-272.

Allen, J., & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the Eighteenth Annual Cognitive Science Society Conference*, 370-375.

Amundson, J. C., & Miller, R. R. (2007). Similarity in spatial origin of information facilitates cue competition and interference. *Learning and motivation*, *38*(2), 155-171.

Antony, J. W., Gobel, E. W., O'Hare, J. K., Reber, P. J., & Paller, K. A. (2012). Cued memory reactivation during sleep influences skill learning. *Nature neuroscience*, *15*(8), 1114-1116.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*(3), 99-102.

Bardovi-Harlig, K. (2000). Tense and Aspect in Second Language Acquisition: Form, Meaning, and Use. *Language Learning: A Journal of Research in Language Studies*, *50*, 1.

Barsky, M. M., Tucker, M. A., & Stickgold, R. (2015). REM sleep enhancement of probabilistic classification learning is sensitive to subsequent interference. *Neurobiology of learning and memory*, *122*, 63-68.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). Mahwah, NJ: Erlbaum.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. *The crosslinguistic study of sentence processing*, *3*, 73-112.

Born, J., Rasch, B., & Gais, S. (2006). Sleep to remember. *Neuroscientist, 12,* 410-424.

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

Breitenstein, C., & Knecht, S. (2002). Development and validation of a language learning model for behavioral and functional-imaging studies. *Journal of Neuroscience Methods, 114*, 173-179.

Breitenstein, C., Zwitserlood, P., de Vries, M., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience, 25,* 493-500.

Brill-Schuetz, K. A., & Morgan-Short, K. (2014). The Role of Procedural Memory in Adult Second Language Acquisition. In *Annual Meeting of the Cognitive Science Society (COGSCI 2014)*,260-265.

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of memory and language*, *32*(1), 76.

Brown, R. M., & Robertson, E. M. (2007). Inducing motor skill improvements with a declarative task. *Nature neuroscience*, *10*(2), 148-149.

Brown, R. M., & Robertson, E. M. (2007). Off-line processing: reciprocal interactions between declarative and procedural memories. *Journal of Neuroscience, 27,* 10468–10475.

Brown, H., Weighall, A., Henderson, L. M., Gaskell, M. G. (2012). Enhanced recognition and recall of new words in 7 and 12 year olds following a period of offline consolidation. *Journal of Experimental Child Psychology, 112,* 56–72.

Bybee, J. (1998b). A functionalist approach to grammar and its evolution. *Evolution of Communication, 2*, 249–278.

Bybee, J. (2007). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *A handbook of cognitive linguistics and SLA* (pp. 216-236). New York and London: Routledge.

Bybee, J. & S. Thompson (2000). Three frequency effects in syntax. *Berkeley Linguistic Society,23*, 65–85.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10130–10134.

Cohen, N. J., and Eichenbaum, H. (1993). Memory, Amnesia, and The Hippocampal System. Cambridge, MA, MIT Press.

Cohen, D. A., & Robertson, E. M. (2011). Preventing interference between different memory tasks. *Nature neuroscience, 14*(8), 953-955.

Cutler, A., & Fodor, J.A. (1979). Semantic focus and sentence comprehension. *Cognition*, 7, 49-59.

Cutler, A., & Norris, D. (1991). Prosody in situations of communication: Salience and segmentation. In *Proceedings, International Congress of Phonetic Sciences* (Vol. 1), 264-270).

Danker-Hopfe, H., Anderer, P., Zeitlhofer, J., Boeck, M., Dorn, H., Gruber, G., … Dorffner, G. (2009). Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of Sleep Research, 18*(1), 74–84.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 364*, 3773–3800.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*(S1), 1-25.

Deliens, G., Leproult, R., Neu, D., & Peigneux, P. (2013). Rapid eye movement and non-rapid eye movement sleep contributions in memory consolidation and resistance to retroactive interference for verbal material. *Sleep*, *36*(12), 1875-1883.

Dewaele, J. M., & Véronique, D. (2001). Gender assignment and gender agreement in advanced French interlanguage: A cross-sectional study. *Bilingualism: Language and Cognition*, *4*(03), 275-297.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11*, 114–126.

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep dependent memory consolidation. *Sleep Medicine Reviews, 13,* 309-321.

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology-General, 121,* 275-277.

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*(2), 169.

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

Durrant, S. J., Cairney, S. A., & Lewis, P. A. (2012). Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum. *Cerebral Cortex*, 1-12.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49,* 1322–1331.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature Reviews Neuroscience, 1,* 41-50.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron, 44*, 109–120.

Eichenbaum, H., & Cohen, N. J. (2001). From conditioning to conscious recollection: Memory systems of the brain. New York: Oxford University Press.

Eidsvag, S. S., Austad, M., Plante, E., & Asbjornsen, A. E. (2015). Input variability facilitates unguided subcategory learning in adults. *Journal of Speech, Language, and Hearing Research, 58*(3), 826-839.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences, 104*, 7723–7728.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in second language acquisition, 24*(02), 143-188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164-194.

Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a complex adaptive system* (Vol. 3). John Wiley & Sons.

Ettlinger, M., Bradlow, A. R., & Wong, P. C. (2014). Variability in the learning of complex morphophonology. *Applied psycholinguistics, 35*(04), 807-831.

Fenn, K. M., Margoliash, D. & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition, 128,* 280-286.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*, 614–616.

Fischer, S., Drosopoulos, S., Tsen, J., & Born, J. (2006). Implicit learning–explicit knowing: A role for sleep in memory system interaction. *Journal of Cognitive Neuroscience, 18,* 311–319.

Fischer, S., Hallschmid, M., Elsner, A. L., & Born, J. (2002). Sleep forms memory for finger skills. *Proceedings of the National Academy of Sciences, 99*(18), 11987-11991.

Fischer, S., Nitschke, M. F., Melchert, U. H., Erdmann, C., & Born, J. (2005). Motor memory consolidation in sleep shapes more effective neuronal representations. *The Journal of neuroscience, 25*(49), 11248-11255.

Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language, 63*(1), 117-130.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218-245.

Gagné, M. H., & Cohen, H. (2016). Interference effects between memory systems in the acquisition of a skill. *Experimental Brain Research*, 1-9.

Gais, S., & Born, J. (2004b). Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences of the USA, 101,* 2140-2144.

Gaskell, 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*(5), 431-436.

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science, 17,* 670–674.

Heine, B., & Kuteva, T. (2007). *The genesis of grammar: A reconstruction* (Vol. 9). Oxford University Press.

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, G. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science, 15*, 674–687.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. Nature Reviews Neuroscience, 11, 523- 532.s

Hennies, N., Lewis, P. A., Durrant, S. J., Cousins, J. N., & Ralph, M. A. L. (2014). Time-but not sleep-dependent consolidation promotes the emergence of cross-modal conceptual representations. *Neuropsychologia*, *63*, 116-123.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology, 10*, 431–436.

Hupbach, A., Gomez, R. L., Bootzin, R. R., & Nadel, L. (2009). Nap-dependent learning in infants. *Developmental Science, 12*, 1007–1012.

Iber, C., Ancoli-Israel, S., Chesson, A., & Quan, S. F. (2007). For the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. 1st ed. Winchester, IL: American Academy of Sleep Medicine*.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. in ba campbell & rm church (eds.), Punishment and aversive behavior (pp. 279-296). *New York: Appleton-Century-Crofts*.

Kastner, S., & Ungerleider, L.G. (2000). Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci*, *23*, 315–341.

Keisler, A., & Shadmehr, R. (2010). A shared resource between declarative memory and motor memory. *The Journal of neuroscience*, *30*(44), 14817-14823.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in second language acquisition*, *20*(04), 543-587.

Knowlton, B. J., & Moody, T. D. (2008). Procedural learning in humans. In. J. H. Bryne (ed.), Learning and memory: A Comprehensive reference (vol. 3): Memory systems, pp. 321–340. Oxford: Academic Press/Elsevier.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, *17*, 3–10.

Lau, H., Alger, S. E., & Fishbein, W. (2011). Relational memory: A daytime nap facilitates the abstraction of general concepts. *PLoS ONE, 6*, e27139.

Lau, H., Tucker, M. A., & Fishbein, W. (2010). Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of Learning and Memory, 93*, 554-560.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron, 36*(6), 1183–1194.

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences, 15*, 343–351.

Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., & Knowlton, B. J. (2004). An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. *Journal of Cognitive Neuroscience, 16*(3), 427-438.

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

MacWhinney, B. (1987). The competition model. *Mechanisms of language acquisition*, 249-308.

MacWhinney, B. (1997). Second language acquisition and the competition model. *Tutorials in bilingualism: Psycholinguistic perspectives*, 113-142.

MacWhinney, B. (1999). The emergence of language from embodiment. In B. MacWhinney (Ed.), *The emergence of language* (pp. 213-256). Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B., & Bates, E. (1989). The crosslinguistic study of sentence processing. Cambridge, England: Cambridge University Press.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of verbal learning and verbal behavior, 23*(2), 127-150.

MacWhinney, B., Pléh, C., & Bates, E. (1985). The development of sentence interpretation in Hungarian. *Cognitive Psychology*, *17*(2), 178-209.

Marshall, U., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences, 11*, 442-450.

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.

Mednick, S., Nakayama, K., & Stickgold, R. (2003). Sleep-dependent learning: a nap is as good as a night. *Nature Neuroscience, 6*, 697-698.

Menon, V., Boyett-Anderson, J. M. & Reiss, A. L. (2005). Maturation of medial temporal lobe response and connectivity during memory encoding. *Cognitive Brain Research, 25,* 379–385.

Mirković, J., & Gaskell, M. G. (2016). Does Sleep Improve Your Grammar? Preferential Consolidation of Arbitrary Components of New Linguistic Knowledge. *PloS one, 11*(4), e0152489.

Mirković, J., Forrest, S. F. and Gaskell, M. G. (2011): Semantic regularities in grammatical categories: Learning grammatical gender in an artificial language. In L. Carlson, C. Holscher, T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 324-329).* Austin, TX: Cognitive Science Society, 324-329.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science, in press*.

Monaghan, P., Chater, N., & Christiansen, M.H. (2005). The differential contribution of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143-182.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259-305.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325.

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition, 17*(1), 56-72.

Munson, B., Swenson, C. L., & Manthei, S. C. (2005). Lexical and phonological organization in children: evidence from repetition tasks. *Journal of Speech, Language, and Hearing Research, 48,* 108-124.

Nemeth, D., Janacsek, K., Londe, Z., Ullman, M.T., Howard, D.V., & Howard, J.H.

(2010). Sleep has no critical role in implicit motor sequence learning in young and old adults. *Experimental Brain Research, 201*(2), 351 - 358.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology, 48*(2), 127-162.

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep Promotes the Extraction of Grammatical Rules. *PLoS ONE, 8*, e65046.

Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PloS one, 2*(4), e341.

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep, 27*, 1255–1273.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 1047-1052). Mahwah, NJ: Lawrence Erlbaum.

Onnis, L., Christiansen, M. H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. In *Proceedings of the 25th annual conference of the cognitive science society* (pp. 886-891). Mahwah, NJ: Lawrence Erlbaum.

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.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L. W., Wamsley, E. J., Tucker, M. A., ... & Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of learning and memory, 92*(3), 327-334.

Peigneux, P., Laureys, S., Fuchs, S., Collette, F., Perrin, F., Reggers, J., Phillips, C., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A., & Maquet, P. (2004). Are spatial memories strengthened in the human hippocampus during slow wave sleep? *Neuron, 44*, 535-545.

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning, 33*, 465–97.

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9,* 534-547.

Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546-550.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377-396.

Rasch, B. & Born, J. (2007). Maintaining Memories by Reactivation. *Current Opinion in Neurobiol. 17*(6), 698-703.

Rasch, B., Buechel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science, 315*, 1426-1429.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 55*, 1339-1362.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. *Current Biology, 14*(3), 208 - 212.

Robertson, E. M. (2009). From creation to consolidation: a novel framework for memory processing. *PLoS Biol*, *7*(1), e1000019.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude, and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfield, and Hernstadt (1991). In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 211–265). Amsterdam/Philadelphia: Benjamins.

Romberg. A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906-914.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Scherag, A., Demuth, L., Rösler, F., Neville, H. J., & Röder, B. (2004). The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects. *Cognition*, *93*(3), 97-108.

Sharwood Smith, M. (1981). Consciousness raising and the second-language learner. *Applied Linguistics, 2*, 159–68.

Smith, R. E., & Hunt, R. R. (2000). The influence of distinctive processing on retrieval-induced forgetting. *Memory & Cognition, 28*, 503–508.

Sokolov, J. L. (1989). The development of role assignment in Hebrew. In B. MacWhinney *& E*. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 158-184). New York: Cambridge University Press.

Song, S., Howard, J.H., & Howard, D.V. (2007). Sleep does not benefit probabilistic motor sequence learning. *The Journal of Neuroscience, 27*(46), 12475 - 12483.

Squire, L.R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*,1380–1386.

Steriade, M. (2003). The corticothalamic system in sleep. *Front Biosci, 8*, 878-899.

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*(7), 1317-1329.

Stickgold, R. (2009). How do I remember? Let me count the ways. *Sleep Medicine Reviews, 13*, 305-308.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience*, *16*, 139-145.

Sweegers, C. C. G., Coleman, G. A., van Poppel, E. A. M., Cox, R., & Talamini, L. M. (2015). Mental schemas hamper memory storage of goal-irrelevant information. *Frontiers in human neuroscience*, *9*, 1–12.

Sweegers, C. C., & Talamini, L. M. (2014). Generalization from episodic memories across time: A route for semantic knowledge acquisition. *Cortex*, *59*, 49-61.

Szmalec, A., Page, M. P. A., Duyck, W. (2012). The development of long-term lexical representations through Hebb-repetition learning. *Journal of Memory and Language, 67*, 342-354.

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

Takashima, A., Petersson, K., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M., . . .

Fernandez, G. (2006). Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 756-761.

Talamini, L. M., Nieuwenhuis, I. L., Takashima, A., & Jensen, O. (2008). Sleep directly following learning benefits consolidation of spatial associative memory. *Learning & Memory*, *15*(4), 233-237.

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., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive psychology*, *79*, 1-39.

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

Tamminen, J., Lambon Ralph, M. A., & Lewis, P. A. (2013). The role of sleep spindles and slow oscillations 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.

Thomas, K. M., Hunt, R. H., Vizueta, N., Sommer, T., Durston, S., Yang, Y., & Worden, M. S. (2004). Evidence of developmental differences in implicit sequence learning: an fMRI study of children and adults. *Journal of cognitive neuroscience*, *16*(8), 1339-1351.

Tomasello, M. (2003). Constructing a language. Cambridge, MA: Harvard University Press.

Tucker, M. A., & Fishbein W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep, 31*, 197-203.

Tucker, M. A., Hirota, Y., Wamsley, E. J., Lau, H., Chaklader, A., & Fishbein, W. (2006). A daytime nap containing solely non-REM sleep enhances declarative but not procedural memory. *Neurobiology of Learning and Memory*, *86*, 241–247.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual review of psychology*, *53*(1), 1-25.

Ullman, M. T. (2001a). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30*(1), 37–69.

Ullman, M. T. (2001b). The neural basis of lexicon and grammar in first and second language: the declarative/ procedural model. *Bilingualism: Language and Cognition*, *4*(1), 105–122.

Ullman, M. T. (2001c). A neurocognitive perspective on language: the declarative/procedural model. *Nature Reviews Neuroscience, 2,* 717–726.

Ullman, M. T. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. *Mind and context in adult second language acquisition*, 141-178.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92,* 231–270.

Ullman, M.T. (2015). The declarative/procedural model: A neurobiologically motivated theory of first and second language. In B. VanPatten and J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed.) (pp. 135-158). New York: Routledge.

Van den Bos, E., Christiansen, M. H., & Misyak, J. B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language*, *67*(4), 507-520.

Van den Bos, E., & Poletiek, F. H. (2015). Learning simple and complex artificial grammars in the presence of a semantic reference field: effects on performance and awareness. *Frontiers in psychology*, *6*, 1–7.

VanPatten, B. (2004). Input processing in second language acquisition. *Processing instruction: Theory, research, and commentary*, 5-31.

Vlach, H. A., Ankowski, A. A., & Sandhofer, C. M. (2012). At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. *Journal of Experimental Psychology: Learning Memory and Cognition, 38,* 246–254.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature, 427,* 352–355.

Wagner, U., Hallschmid, M., Verleger R., & Born, J. (2003). Signs of REM sleep dependent enhancement of implicit face memory: a repetition priming study. *Biol Psychol, 62*,197–210.

Walker, M., Brakefield, T., Morgan, A., Hobson, J., & Stickgold, R. (2002). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron, 35*, 205-211.

Walker, M., Brakefield, T., Seidman, J., Morgan, A., Hobson, J., & Stickgold, R. (2003). Sleep and the time course of motor skill learning. *Learning & Memory, 10*, 275-284.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron, 44,* 121–133.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. *Nature Reviews Neuroscience, 11*, 218.

Wamsley, E. J., & Stickgold, R. (2011). Memory, sleep, and dreaming: Experiencing consolidation. *Sleep Medicine Clinics, 6*(1), 97–108.

Werchan, D. M., & Gomez, R. L. (2014). Wakefulness (not sleep) promotes generalization of word learning in 2.5-year-old children. *Child Development*, 1-8.

Wilhelm, I., Metzkow-Mészàros, M., Knapp, S., & Born, J. (2012). Sleep-dependent consolidation of procedural motor memories in children and adults: The pre-sleep level of performance matters. *Developmental science*, *15*(4), 506-515.

Wilhelm, I., Rose, M., Imhof, K. I., Rasch, B., Büechel, C., & Born, J. (2012). The sleeping child outplays the adult's capacity to convert implicit into explicit knowledge. *Nature Neuroscience, 16,* 391-393.

Williams, S. E., & Horst, J. S. (2014). Goodnight Book: Sleep Consolidation Improves Word Learning via Storybooks, *Frontiers in Developmental Psychology, 5*, 1-12.

Yordanova, J., Kolev, V., Wagner, U., & Verleger, R. (2010). Differential associations of early and late-night sleep with functional brain states promoting insight to abstract task regularity. *PLoS One, 5,* e9442.

Yordanova, J., Kolev, V., Wagner, U., Born, J., & Verleger, R. (2012). Increased alpha (8–12 Hz) activity during slow wave sleep as a marker for the transition from implicit knowledge to explicit insight. *J Cogn Neurosci, 24*, 119–132.