Vocabulary Learning and Regularity Extraction: Temporal Dynamics of Consolidation and Associations with Slow-wave Sleep and Sleep Spindles

Eva Kimel[1,2*], Ilana S. Hairston[1,3], Dafna Ben-Zion[1], Yekete Akal[3], Anat Prior[1], M. Gareth Gaskell[2], and Tali Bitan[1,4]

[1] University of Haifa, Haifa, Israel, [2] University of York, York, UK, [3] Tel Hai Academic College, Tel-Hai, Israel, [4] University of Toronto, Toronto, ON, Canada
* Corresponding author

This research was funded by Israeli Science Foundation grant no. 1052/16 to T. Bitan. E. Kimel was supported by the Israel National Postdoctoral Award for Advancing Women in Science.

We would like to thank Adi Marinberg, Mona Blyer, Haya Hajyahya, Amit Green, Oren Levin, and Rabab Fadul for their help with participant recruitment and data collection, and to Dan Denis and Scott Cairney for their valuable advice.

Vocabulary Learning and Regularity Extraction: Temporal Dynamics of Consolidation and Associations with Slow-wave Sleep and Sleep Spindles

Abstract: Fast sleep spindles and slow-wave sleep (SWS) have been linked to memory consolidation, however, their associations with learning and longer term retention of different aspects of language remain unclear. We investigated the temporal dynamics of consolidation of vocabulary and grammar, and their links with these sleep metrics. Young adult participants were trained in the evening on an artificial language that used plural inflections with an underlying morpho-phonological regularity that was not taught explicitly. Some of the words were presented frequently and others infrequently. Polysomnographic measures were collected during the night following learning; participants were tested on the vocabulary, trained inflections, and generalisation to untrained words at four time points across nine days.

Accuracy on the vocabulary test improved across the first night following learning, and the change was positively associated with SWS duration. Memory for infrequent words declined towards Day 9, but greater spindle density during the first night was associated with a smaller decline. Although mean group accuracy on trained inflections did not significantly change overnight, individually, the change was negatively correlated with spindle density. Generalisation accuracy showed no change across time and no correlations with sleep characteristics. Overall, the results demonstrate that vocabulary and grammar learning have different temporal dynamics of consolidation and distinct patterns of association with sleep metrics. The findings suggest a protective role of spindles for long-term retention of memory, particularly of weakly encoded items, and emphasise the need to dissociate the benefits of SWS from those of spindles.

Introduction

Language learning is a continuous process that begins before birth and continues throughout life (Brysbaert et al., 2016), relying on both encoding and offline consolidation. The process of learning, or forming new memories, does not end with exposure, but rather continues to undergo additional changes in the following hours, days, weeks and longer (e.g., Gaskell, 2024). During this consolidation process, the new memory trace becomes more stable and is integrated with previously existing memories (Diekelmann & Born, 2010; Gaskell & Dumay, 2003; Stickgold & Walker, 2005). Consolidation can also facilitate the extraction of patterns, regularities, or implicit rules that in turn enable generalisation: applying knowledge from previously learned information when processing new input (Fenn et al., 2003; Tamminen et al., 2012).

A recent systematic review and meta-analysis on word learning confirmed that various aspects of memory for novel words benefit from sleep (Schimke et al., 2021). Sleep was shown to benefit both recall and recognition of novel word forms, as compared to a time period that does not contain sleep (Mirković & Gaskell, 2016; Tamminen et al., 2010). Sleep also supports the integration of novel word forms into the existing lexicon, as first demonstrated in a pioneering study by Gaskell and Dumay (2003). In this study, participants were taught novel word forms that were phonological neighbours of existing words (e.g., *cathedruke* - *cathedral*). They found inhibition in the access to existing words due to competition with the new word forms, but this was evident only in the delayed (one week post-training) test and not immediately after training, suggesting that offline consolidation is required for novel words to become fully integrated into the mental lexicon. The benefit of a delay that includes sleep for lexical integration has since been replicated in several studies (Davis et al., 2009; Davis & Gaskell, 2009; Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008; Van Der Ven et al., 2015).

In contrast to vocabulary learning, the contribution of sleep to offline extraction of regularities in language learning is less clear. Some previous studies have demonstrated sleep-dependent enhancement in inferring covert relationships between non-linguistic stimuli (e.g., Durrant et al., 2011, 2013; Lewis & Durrant, 2011; Wagner et al., 2004, Ellenbogen et al., 2007), leading to the suggestion that sleep may also benefit rule extraction and generalisation in language stimuli (e.g., Batterink et al., 2014; Batterink & Paller, 2017). For example, in a study in which participants learned new affixes, learning effects were assessed both with speed of oral repetition and with a definition selection task. When the learned affixes were attached to untrained stems in order to assess generalisation, the advantage in the speed of oral repetition was only evident 2 days after training, supporting the contribution of offline consolidation to generalisation. Interestingly, in the definition selection task, participants showed evidence of generalisation immediately after learning, emphasising task-dependence (Tamminen et al., 2010; for a review see Cordi & Rasch, 2021; Palma & Titone, 2021).

However, the role of sleep in generalisation in language was questioned in a number of studies. For example, Tamminen and colleagues (2020) introduced adult participants to a novel orthography. Participants were able to generalise this knowledge to unfamiliar words, even when they were deprived of sleep the night after learning. Moreover, Mirković & Gaskell (2016) did not find a benefit of a post-learning nap for language regularity extraction, in contrast to a sleep benefit for vocabulary learning found in the same study. A similar result was reported by Ben-Zion and colleagues (Ben-Zion et al., 2022): They exposed participants to artificial novel words and their plural endings, where the plural suffixes were implicitly determined by the stem endings. A group that trained in the evening and slept immediately after training was compared to a group that trained in the morning and thus slept ~12 hrs after training. While inflection of trained words, which can rely on item-specific knowledge, differed between the groups 12-hours post-learning, the generalisation to untrained items, which reflects the extraction of regularities improved after 24 hours, but showed no group difference in performance. In summary, sleep plays a central role in vocabulary learning, but its impact on regularity extraction and generalisation is not as straightforward. Importantly, most studies have looked at either vocabulary learning or generalisation (though see Mirković & Gaskell, 2016), making it harder to directly compare learning of these two aspects. We address this limitation in the current study by examining both vocabulary and grammar learning together.

The Complementary Learning Systems framework (CLS; McClelland et al., 1995; O'Reilly et al., 2014) proposes a model for the involvement of sleep in memory formation. It suggests that during exposure, an initial episodic, context-rich encoding is formed, supported by the hippocampus; then, during sleep, the memories are integrated into cortical networks and become less dependent on the hippocampus (Klinzing et al., 2019; Kumaran et al., 2016). Davis and Gaskell (2009) were the first to apply this model to language learning by reviewing behavioral and brain imaging studies that together support a two stage learning process as described by the CLS. First, rapid word acquisition supported by the medial temporal lobe and the hippocampus, followed by consolidation that involves offline neocortical learning, the evidence to which emerges following a night of sleep (Gaskell, 2024).

Several sleep metrics have been linked to sleep-dependent memory consolidation in language, with two key measures being slow-wave sleep duration (Tamminen et al., 2010) and sleep-spindle density (Tamminen et al., 2013; Tham et al., 2015). Slow-wave sleep (SWS) is the sleep stage that is characterised by the most synchronised and low frequency neural activity; Sleep spindles are brief bursts of neural activity lasting .5–3 seconds within a specific frequency range (12-16 Hz; Ng et al., 2024), and their role in memory consolidation was confirmed by a recent meta-analysis (Kumral et al., 2023) and was linked to memory replay (e.g., Cairney et al., 2018). Both SWS and sleep spindles are thought to support the transfer of information from relying on the hippocampus to relying primarily on the neocortex (Klinzing et al., 2019) with a co-occurrence of spindles and slow oscillations (typical of slow-wave sleep) identified as a predictor of memory consolidation across a sleep period (Denis & Cairney, 2023; Staresina, 2024).

More specifically, for language learning, SWS duration has been positively correlated with vocabulary acquisition, as measured by recognition speed of newly learned words (Tamminen, 2010), and by paradigms that assess automatic access to word meanings (Tham et al., 2015). However, there is no evidence regarding a direct correlation of SWS duration and regularity extraction in language. SWS duration was positively associated with learning of implicit restrictions on phoneme position within syllable sequences, but not with generalisation of these constraints to novel sequences (Gaskell et al., 2014). Similarly, others did not find correlations between SWS duration and generalisation of linguistic regularities (Batterink et al., 2014; Batterink & Paller, 2017). For example, Batterink and colleagues (2014) introduced an implicit rule where novel articles predicted noun animacy. Participants showed slower responses for untrained phrases that violated the rule, thus exhibiting generalisation, but their sensitivity to the rule was not correlated with SWS duration.

Sleep-spindle activity was found to be correlated with proper-name learning (Clemens et al., 2005), with the integration of novel words into the lexicon (Tamminen et al., 2010, 2013; Tham et al., 2015), and with overnight change in cued-recall accuracy for novel words (Weighall et al., 2017). To our knowledge only one study, by Batterink and Paller (2017), tested the association of sleep spindles with regularity extraction, and they did not find a significant association.

To summarise, vocabulary learning benefits from sleep and these benefits are associated with SWS duration and sleep-spindle density: two aspects of sleep heavily implicated in memory consolidation. In contrast, it is unclear whether sleep benefits regularity extraction, and the few studies that tested associations of regularity extraction and sleep characteristics did not find a correlation with either SWS or spindles. Therefore, the aim of the present study is to systematically investigate the associations of vocabulary learning and regularity extraction with both SWS duration and spindle density in the same cohort of participants.

Moreover, we will also test longer-term memory, 8 days after initial training, based on reports of some improvements in the integration of novel words into the mental lexicon becoming evident only a week after the initial training (Clay et al., 2007; James et al., 2019; Tamminen & Gaskell, 2012; though see Tamminen et al., 2013). Importantly, previous research has only studied the correlations between sleep metrics and immediate post-learning or post-sleep performance, but the associations of SWS and spindles with longer-term consolidation have not been tested. Therefore, a key strength and novel aspect of our study lies in exploring the connection between post-learning sleep characteristics and longer-term retention.

Finally, we manipulated frequency of exposure - in light of the reports on preferential sleep-dependent consolidation of weaker memories (Denis et al., 2020, 2021; Diekelmann et al., 2009; Drosopoulos et al., 2007; Schapiro et al., 2017), we wanted to test this effect on word learning.

To this end we adapted an artificial language used in previous studies (Ben Zion et al., 2019; Ben-Zion et al., 2022) which have shown evidence for learning of implicit morpho-phonological regularities and facilitation by sleep for trained words (i.e., stem+suffix), though not for generalisation. We added a direct assessment of vocabulary learning and included four time points in our study: 1) in the evening - immediately after training, 2) the next morning, after a night of sleep (~12 hrs post-training), 3) the following morning, a day after (~36 hrs post-training) and 4) six days after session 3. The training comprised novel vocabulary and plural forms, with implicit morpho-phonological regularity underlying the plural form suffixes.

We tested the associations of SWS duration and sleep spindles with the change in accuracy for vocabulary, trained plural inflections and generalisation across the four time points, which spanned nine days (generalisation was only measured across three days). We predicted a positive association of the sleep metrics with the change in performance over time for vocabulary and plural inflections, with stronger association between sleep characteristics and memory for infrequent words as compared to frequent words. For grammar learning, based on previous findings of no correlations, we hypothesised that any positive associations might only become evident on Day 3, rather than the day following initial learning.

Methods

Participants

We analysed the data of 29 participants (F = 19), whose mean age was 25.83 ± 3.28 years. Participants were recruited in and around the campus of Tel-Hai Academic College (Kiryat Shmone, Israel); They could choose whether to be paid or receive research participation credit. We invited participants whose native language was Hebrew to minimise linguistic background variability, who had normal or corrected-to-normal vision. Exclusion criteria included: diagnosed hearing deficits, neurological or psychiatric diagnoses, habitual use of medications affecting sleep, habitual daytime napping, regular smoking, travelling across time zones in the past two weeks, and pregnancy. We based our power analysis on the study by Tamminen et al. (2010), which assessed correlations between behavioural measures and both SWS duration and sleep-spindle density. Using Fisher's z-transformation, a power of .80 and a two-sided alpha < .05 yielded an estimated sample size of 27 participants. To account for potential data loss, we decided to recruit 30 participants.

Participants then filled screening questionnaires and participated in an introductory session via videoconference. Sleep disorders were ruled out by the Mini Sleep Questionnaire (MSQ; Natale et al., 2014). To increase the likelihood of participants falling asleep in the sleep lab, we did not invite those classified as “evening type” based on the Hebrew version of the Morningness-Eveningness Questionnaire (MEQ; Horne & Ostberg, 1976). We did not invite participants who were diagnosed with any learning or communication difficulty, nor individuals who scored ≥ 51 on a ADHD questionnaire (Zohar & Konfortes, 2010) so that not to introduce heterogeneity that might be relevant to language processing (Rucklidge & Tannock, 2002) or sleep patterns (Becker, 2020; Konofal et al., 2010). We stopped recruitment after testing 30 adults (18-45 years old). Data of one participant were excluded from the analysis due to poor sleep quality (18 awakenings of at least one minute, only 2 REM periods), resulting in 29 usable datasets.

Experimental Procedure

The study protocol was approved by the Ethics committees of Tel-Hai college (approval 10/2019-4), and the University of Haifa (approval 378/19). The 30 participants who passed initial screening were invited to take part in the study that spanned nine days (Fig. 1). On the three nights before they began the experiment, participants were requested to go to bed no later than midnight and allow a sleep of 7-8 hours, and not to nap during the day. They filled a sleep log reporting their sleeping times and were given a smart watch to verify their sleep times on the days before the experiment. On Day 1, participants were instructed to refrain from consuming alcohol or other psychoactive substances, and only consume caffeine before 2pm. In the sleep laboratory, participants went to bed by 22:45.

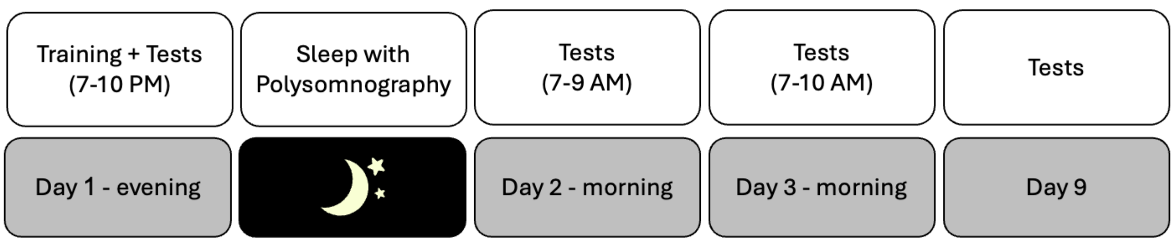

Figure 1. Study overview.

Stimuli, tasks and behavioural data preprocessing

The design was based on paradigms used in our previous studies (Ben Zion et al., 2019; Ben-Zion et al., 2022, 2023; Nevat et al., 2017, 2018), and included learning singular and plural words in an artificial language. The first session included training and testing, whereas all other sessions included tests only (Fig. 1).

Participants learned 36 new words which were paired with existing objects. Each training trial included an auditory presentation of the singular form (i.e., stem), the plural form, and an image depicting its referent. For example, participants heard *refoz*, then *refozan*, and were presented with an image of an apple. All stems had a CVCVC structure; Plural forms consisted of the stem and one of three VC suffixes (*an, esh, ur*). The last two phonemes of the stem determined the suffix, such that certain stem endings were matched to a certain suffix. For example, the plural suffix for stems ending with either */oz/* or */ap/*, was *an.* The mappings between the stem ending and the plural suffix that underlied the plural form rules were arbitrary.

The training set included 30 words that followed these rules: 10 stems per each plural suffix. Frequency of word presentation varied such that 18 words were presented nine times each (i.e., frequent) and 12 words were presented three times each (i.e., infrequent) during training. In addition, six "irregular" words that violated the rules described above were presented. These were included to increase similarity to natural languages, and were not included in the analyses.

Day 1 session consisted of: <u>Exposure</u> - On each trial participants were auditorily presented with the singular and plural form of one novel word, and with its image referent (for 1 second). They were requested to remember the singular form-referent mapping and to repeat out loud the plural form (with a timeout of five seconds). The trials were self-paced; trial order was random. Each of the 36 words was presented once. <u>Test of trained inflections</u> (baseline) - On each trial participants were auditorily presented with a singular form and a plural form and were required to indicate whether the plural form corresponded to the singular form by pressing key *1* or *2* on the keyboard (with a timeout of three seconds). Each trained word appeared twice in the test, once with its correct plural form and once with an incorrect plural form. For the incorrect plural form, a wrong combination with one of the other two trained plural endings was used in an alternating manner (e.g., if the correct ending was *an*, and *esh* was used in Day 1 test, then *ur* will be used in Day 2 test, etc.). Trial order was random. <u>Training</u> - On each trial participants were auditorily presented with a singular form and saw its image referent. They were then asked to say the plural form out loud (with a timeout of three seconds), followed by an auditory presentation of the correct plural form. Training comprised three blocks with a short break between them; frequent items were presented three times during each block, and infrequent items were presented once. The trials were self-paced; trial order was random. <u>Test of trained inflections</u> (post-training) - as described above. <u>Vocabulary test</u> - On each trial participants were auditorily presented with a singular form and saw an image referent, and were asked to indicate whether the image corresponds to the singular form, by pressing key *1* or *2* on the keyboard (with a timeout of three seconds). Each trained stem appeared twice in the test, once with its correct corresponding image and once with an image that is the referent of a different word in the experiment. Distractor images were chosen randomly such that in each test each image appeared in one trial as the correct image (i.e., with the stem that it was paired with during training), and in one trial as the incorrect image (i.e., with a stem that it was not paired with during training). Trial order was random. Generalisation test - On each trial participants heard a singular word form that was not included in the training set but had one of the endings of trained stems (e.g. */oz/* or */ap/*), and were asked to produce the plural form of that word (with a timeout of three seconds). Thirty novel words were presented in each test, five for each of the six phonological stem cues. The trials were self-paced; trial order was random. For more details please refer to Ben Zion et al. (2019).

Day 2 and Day 3 sessions included the three tests (trained inflections, vocabulary, generalisation); Day 9 session included tests of trained inflections and vocabulary only (due to technical reasons generalisation was not tested). The experiment was conducted using Matlab (Inc, 2022).

Analyses were performed with accuracy as the dependent variable, defined as selection of the correct option in the vocabulary and trained inflection tasks, and verbally producing the plural form with the correct suffix (even if the pronunciation of the stem that was part of the plural form was compromised) in the generalisation task. The independent variables for vocabulary and trained inflection tasks were word frequency during training (frequent vs. infrequent) and post-training timepoints (Day 1, Day 2, Day 3, Day 9). The independent variable for the generalisation tasks was post-training timepoints (Day 1, Day 2, Day 3). We also analysed the reaction times (RTs) of correct responses in order to verify that positive changes in accuracy did not result in significantly slower responses.

Polysomnography data collection and preprocessing

Polysomnography (PSG) data collection was performed at the Research Institute of Applied Chronobiology at Tel Hai Academic College. PSG measurements were acquired using SOMNOscreen™ (Somnomedics, Germany). The montage included seven electroencephalogram (EEG) channels (F3, F4, C3, C4, Cz - as reference, A1, A2), bilateral electrooculogram (EOG), submental electromyogram (EMG), and electrocardiogram (ECG). Signals were digitised at 256 Hz, with low- and high-frequency filter settings at 0.2-35 Hz for EEG, 0.2-10 Hz for EOG and 10-35 Hz for EMG, and a 50 Hz notch filter was applied to further minimize electrical noise. A trained sleep technician collected and scored the data in accordance with the American Academy of Sleep Medicine guidelines (AASM; Iber, 2007). All reported duration measures are thus based on this scoring. Prior to spindle detection, a research assistant conducted an additional examination of the data to exclude noisy periods that did not span entire epochs. Using the MNE-Python package, the data were down-sampled to 128 Hz, re-referenced to mastoid (A1, A2) average, band-pass filtered to .3-30 Hz. Independent Component Analysis (ICA) was conducted using the MNE package (Gramfort, 2013), and the signal was reconstructed after removing the three main components: cardiac interference, a salient noise/distortion, and eye movements.

The duration of SWS (Stage 3) was extracted from the scored data, and *SWS duration*, which is the summed duration of all SWS periods in minutes, was used as the measure for each participant. Fast sleep spindles (12-16 Hz, duration .5-3 seconds; Mölle et al., 2011; Ng et al., 2024) were detected in all Stage 2 and Stage 3 epochs (henceforth NREM; Cairney et al., 2018; Leach et al., 2024; Tamminen et al., 2020) using the YASA toolbox for python (relative power = .1, correlation with spindle freq. = .45, amplitude in filtered signal = 2.5; Vallat & Walker, 2021). Average spindle density across the four EEG electrodes (Cairney et al., 2018; although see Mölle et al., 2011) was then used as a single spindle measure for each participant. In order to verify the validity of averaging across the four electrodes, we conducted a principal component analysis (PCA) on the spindle data across the four electrodes. The first component had a very strong correlation ($r = .99$) with the mean of the four electrodes, its loadings were similar across all four electrodes (range .68 - .86) and it captured .62 of the total variance, supporting the decision to use the average as a single measure.

Statistical analysis

Statistical analysis was performed in R (R Core Team, 2021), and plots were produced in python (Van Rossum & Drake, 2009). Linear mixed-effects models were constructed using the *lme4* package (Bates et al., 2015) in *R*. For each of the three tasks: vocabulary, trained inflections, and generalisation, we first defined a full model including all predictors as fixed effects, by-participant intercepts, and by-participant slopes for time points, and for word frequency where applicable (i.e., for vocabulary and trained inflections). By-participant random effects were added to the models in order to account for variability that is not directly related to the effects of interest.

The dependent variable was accuracy: a binary outcome per-trial. We used the *Buildmer* function to find the maximal model that can still converge and which includes all fixed factors. For that model, we ran the *glmer* function from the *lme4* package. We also analysed the reaction times (RTs) of correct responses using the same procedure in order to verify that positive changes in accuracy did not result in significantly slower responses.

For accuracy in each task, we first ran a behaviour-only model (with time point as a predictor, and with word frequency as a predictor for vocabulary and trained inflections) and then ran a model with the behavioural predictors and the examined sleep characteristic: centralised spindle density or SWS duration.

We configured factor coding using *code_diff* from the R package *codingMatrices* (https://CRAN.R-project.org/package=codingMatrices) resulting in contrasts that are the successive differences of the means, $\mu i{+}i - \mu i$. This resulted in the following contrasts for the time points: *Day 2 - Day 1; Day 3 - Day 2; Day 9 - Day 3.* For frequency, this definition resulted in one contrast. For follow-up analyses on interactions within the models, we applied Holm-Bonferroni correction according to the number of analyses for that model (e.g., follow-up analyses within frequent and infrequent words were corrected to 2 comparisons).

Pearson's r was used to report correlation coefficients between tasks (vocabulary, trained inflections, and generalisation) in both single time-points and the intervals of interest in order to examine commonalities between tasks in their consolidation trajectories (see Ben-Zion et al. 2023). Significance was tested using the Holm-Bonferroni procedure (Holm, 1979) according to the number of tests involving the same constructs (e.g., the correlation of generalisation and vocabulary tasks were assessed at 3 time points and across 2 time intervals, thus we corrected for 5 comparisons).

Results

According to manual sleep staging using 30-second epochs, the mean(SD) for total sleep time was 7.67 (.59) hours, with number of sleep cycles = 4.14 (1.04), Stage 1 = 2.84% (1.67%), Stage 2 = 56.30% (7.21%), Stage 3 = 21.48% (4.57%), REM = 19.38% (5.27%).

The first sections report analyses of accuracy and RT for each of the three tasks: vocabulary, trained inflections, and generalisation. The following sections add to the accuracy models the assessed sleep measures: SWS duration and sleep spindles density, separately.

Vocabulary

At all post-training time points accuracy was significantly above chance ($p = 10^{-15}$; Fig. 2) for both frequent and infrequent words. The mixed-effects model (Table S1A) revealed a significant main effect of word frequency ($z = 3.9$, $p = 10^{-4}$), with higher accuracy for frequent words. There was a significant increase in accuracy over the first night after learning ($z = 2.6$, $p = .010$), and a significant decline in accuracy in the third interval: Day 3-morning to Day 9 ($z = 2.1$, $p = .033$). The decline in the third interval was greater for infrequent words than for frequent words (*frequency x time_point* interaction: $z = 2.2$, $p = .028$). In fact, a significant decline over the third interval was found only for infrequent words ($z = 2.88$, $p = .004$; statistically significant according to the Holm-Bonferroni method with $\alpha = .01$), but not for frequent words ($z = .053$ $p = .958$), as revealed by a follow-up analysis with the model *accuracy ~ 1 + time_point + (1 | participant)* separately for frequent and infrequent words. There were no other significant effects or interactions. In summary, accuracy for frequent words was higher than for infrequent ones, overall accuracy increased across the first night, and accuracy of infrequent items only declined over longer intervals.

We analysed the reaction times (RTs) of correct responses using the same mixed model structure as for accuracy in order to verify that positive changes in accuracy did not result in significantly slower responses. The model revealed a significant effect of frequency ($t = 2.13$, $p = .033$) with RTs being faster for frequent vs. infrequent words. The model also revealed a significant decrease in RTs in the second and third intervals ($t = 4.45$, $p = 10^{-4}$; $t = 2.62$, $p = .014$ respectively). There were no other significant effects or interactions (Fig. S1; Table S1B).

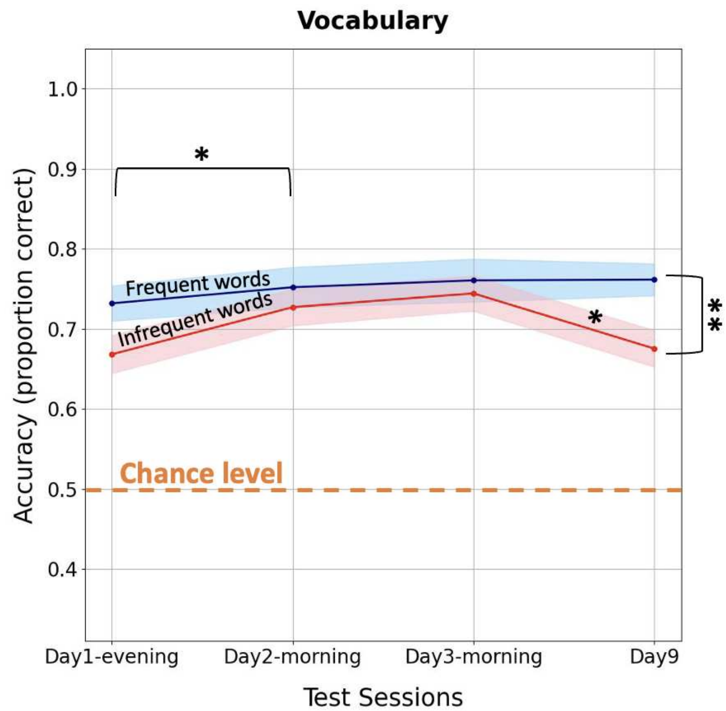

Figure 2. Average performance for the vocabulary task in the four time points; shaded areas denote standard errors. * p < .05; ** p < .0005.

Trained inflections

For all post-training time points mean accuracy was significantly above chance ($p = 10^{-13}$; Fig. 3) for both frequent and infrequent words. The mixed-effects model revealed a significant main effect of word frequency ($z = 4.3$, $p = 10^{-4}$), with higher accuracy for frequent words. In addition, there was a significant change in performance over the second interval ($z = 2.5$, $p = .014$) with performance improving between Day 2-morning and Day 3-morning (Fig. 3). There were no other significant effects or interactions (Table S1C). In summary, accuracy for frequent words was higher than for infrequent ones, and overall accuracy increased across the second interval.

We analysed the reaction times (RTs) of correct responses in order to verify that positive changes in accuracy did not result in significantly slower responses: The model revealed a significant effect of frequency ($t = 5.73$, $p = 10^{-8}$) with RTs being faster for frequent vs. infrequent words. There were no other significant effects or interactions (Fig. S2; Table S1D).

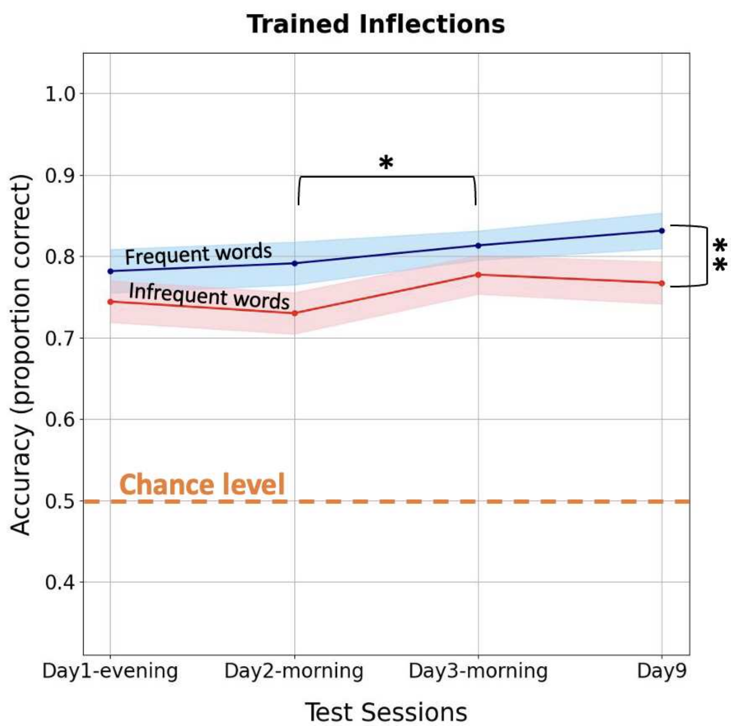

Figure 3. Average performance for the trained inflections task in the four time points; shaded areas denote standard errors. * p < .05; ** p < .0005.

<u>Generalisation</u>

To determine whether production of the correct plural form was above chance, we first assessed the proportion of participant productions that ended with a suffix other than *an, esh, ur* - the three suffixes introduced in the study. The proportion of such errors was very low (in the Day 1 post-training test: .04 (.07), Day 2: .04 (.06), Day 3: .02 (.03)). We thus used a chance level of ⅓, as there were essentially three plural suffixes that participants selected from. For all post-training assessments, mean group accuracy was significantly above chance (p = 10^{-5} ; Fig. 4). The mixed-effect model revealed no significant effects, suggesting that performance did not change significantly between the three time points (Table S1E; Fig. 4).

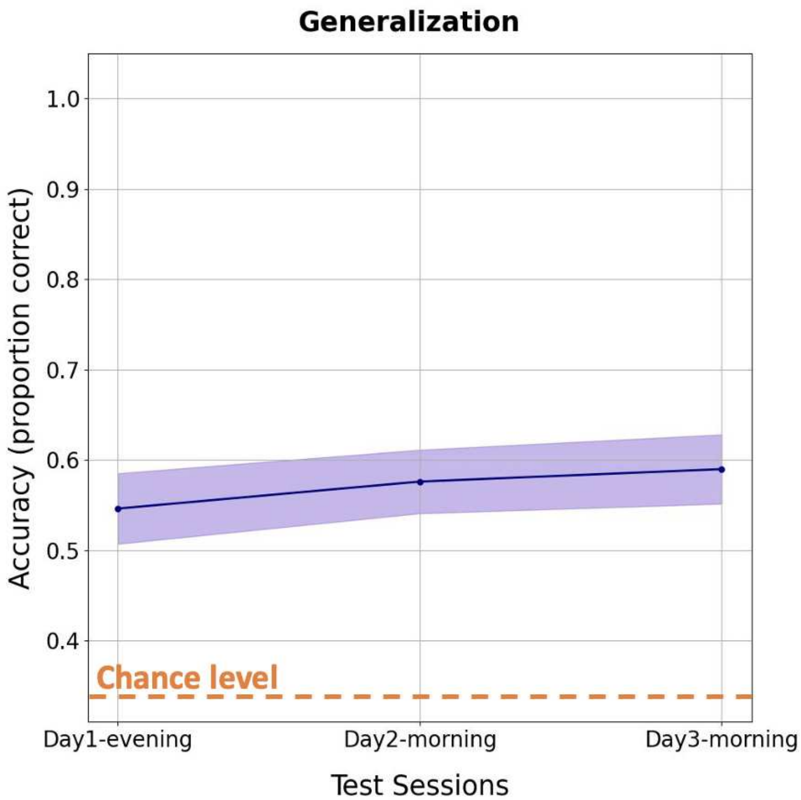

Figure 4. Average performance for the generalisation task in the four time points; shaded area denotes standard errors.

Performance on the trained inflections task and on the generalisation task was significantly correlated at all time points (Table 1). Vocabulary performance significantly correlated with the inflections tasks starting from the second test point (Day 2). In contrast, the change across the intervals did not correlate between the tasks (Table 2).

Table 1. Correlation between performance in the three tasks at each testing session. Pearson correlation coefficients and the corresponding p-values (uncorrected) are presented; Statistical significance was assessed using the Holm-Bonferroni method for multiple comparisons. * p after correction < .05; ** p after correction < .001.

Testing session:	**Day 1**	**Day 2**	**Day 3**	**Day 9**
Vocabulary and Trained inflections	r = .31, p = .098	r = .40, p = .031	r = .42, p = .022	r = .23, p = .237
Vocabulary and Generalisation	r = .34, p = .072	**r = .55, p = .002 ***	**r = .47, p = .010 ***	_
Trained inflections and Generalisation	**r = .65, p = .0001 ****	**r = .65, p = .0001 ****	**r = .63, p = .0002 ****	_

Table 2. Correlation between the change in performance in the three tasks across the intervals between testing sessions. Pearson correlation coefficients and the corresponding p-values (uncorrected) are presented.

Interval:	**Day 1 to Day 2**	**Day 2 to Day 3**	**Day 3 to Day 9**
Vocabulary and Trained inflections	r = -.18, p = .356	r = .12, p = .534	r = -.09, p = .633
Vocabulary and Generalisation	r = -.08, p = .687	r = .27, p = .155	_
Trained inflections and Generalisation	r = .07, p = .733	r = -.04, p = .857	_

The association of language learning with the duration of SWS

We tested whether change in accuracy in each of the three tasks (vocabulary, trained inflections, generalisation) was predicted by the duration of SWS (in minutes), by adding SWS duration (in minutes) to the predictors used in the behaviour-only model: namely, frequency and time point. Total sleep time was also included in the model as a control variable that allowed testing for associations with SWS beyond total sleep time.

For vocabulary, we found a positive association of SWS duration with the change in performance over the first interval: Day 1-evening (immediately post-training) to Day 2-morning (z = 2.39, p = .017; Fig. 5), and there were no other significant effects or interactions (Table S2A) beyond those reported in the behaviour-only model. In order to verify that the association between the duration of SWS and the change in performance over the first interval was not due to outliers, we excluded any outliers above/below 2.5 SDs. This exclusion resulted in removing the highest value of SWS duration (i.e., 157.5 minutes, *Z-score* = 2.80), we tested the correlation without this value, and it remained significant (*Pearson r* = .50, p = .007).

For trained inflections, there were no significant associations with SWS duration (Table S2B).

For generalisation, there were no significant associations with SWS duration. The model revealed a significant positive interaction between the first interval (Day 2 vs. Day 1) and total sleep duration, which was used as a control variable in the model (z = 2.5, p = .012). No other factors or interactions were found to be significant (Table S2C).

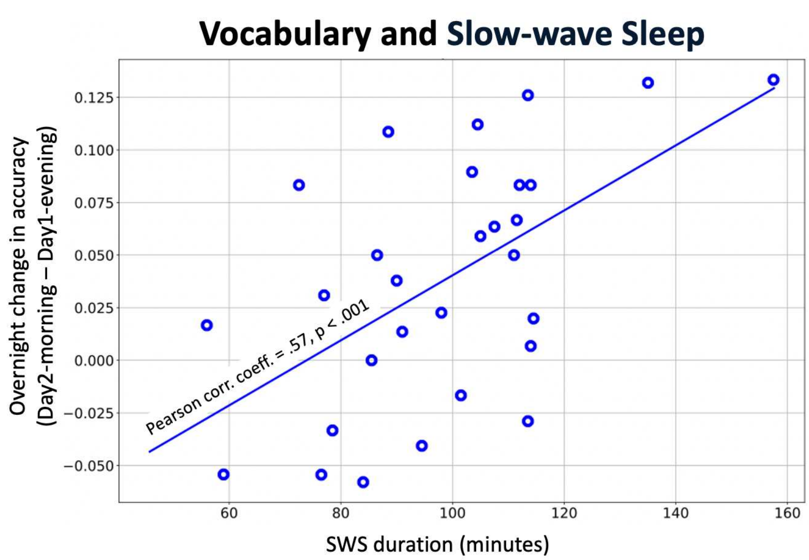

Figure 5. A positive association of SWS duration with the change in accuracy between the immediate evening test (Day 1) and the following morning (Day 2) in the vocabulary task. Each dot denotes the data of one participant, the line is a group linear regression line.

The association of language learning with fast sleep spindles

To test whether spindles during NREM predicted performance in each of the three tasks, we included fast spindle density as a predictor in the models predicting performance, in addition to word frequency and time-point.

For <u>vocabulary</u>, there was a significant *frequency x time_point x spindle density* interaction (z = 2.25, p = .025; Standardised coefficient = .37, 95% CI [.05, .70] indicating a moderate effect) for the third interval (Day 3-morning to Day 9). Two follow-up analyses on this interval were conducted with the model *accuracy ~ 1 + time_point x spindle_density + (1 | participant)* separately for frequent and infrequent words. As in the behaviour-only model, a significant decline in performance was found for infrequent words (z = 2.83, p = .005; statistically significant according to the Holm-Bonferroni correction with α = .05) but not for frequent words (z = .025 p = .980). In addition, a significant positive *time_point x spindle_density* interaction was found for infrequent words (z = 2.27, p = .023) but not for frequent words (z = .77, p = .439): Spindle density positively correlated with the change in accuracy across the third interval for infrequent words only (Fig. 6) such that higher spindle density was associated with smaller forgetting from Day 3 to Day 9. No other significant effects or interactions were found beyond those already reported in the behaviour-only model (Table S3A).

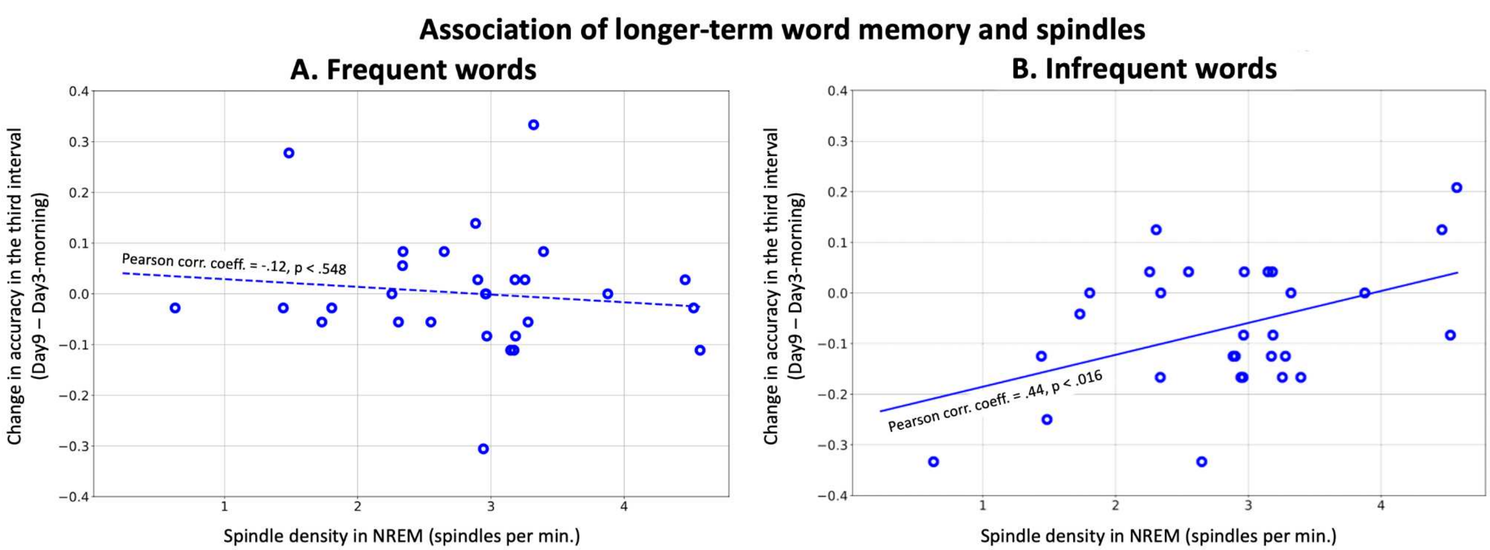

Figure 6. A positive association of the density of sleep spindles in NREM with the change in accuracy in the vocabulary task for infrequent (B), but not frequent (A), words in the third interval: Day 9 vs. Day 3-morning.
Each dot denotes the data of one participant, the line is a group linear regression line.

For trained inflections, we found a negative association between the spindle density and the change in performance in the first interval[1]: post-training Day 1 to Day 2 (z = 2.50, p = .013; Fig. 7; Table S3B), and there were no additional effects or interactions beyond what was reported in the behaviour-only model (Table S3B). To test if this negative association is due to a negative correlation between performance at the end of training and overnight change we tested the correlation between the two. Participants who performed better at the end of training (pre-sleep) showed a smaller overnight improvement (*Pearson* r = -.54, p = .004). However, there was no association between spindle density and pre-sleep performance for the trained inflections (*Pearson* r = .090, p = .642).

For generalisation, there were no significant effects or interactions with spindles (Table S3C).

[1] In response to a reviewer request, we tested the negative correlation between spindle density and the residuals of the morning scores, after regressing out the immediate scores. The correlation remained significant: Pearson's ρ = -.46, p < .013.

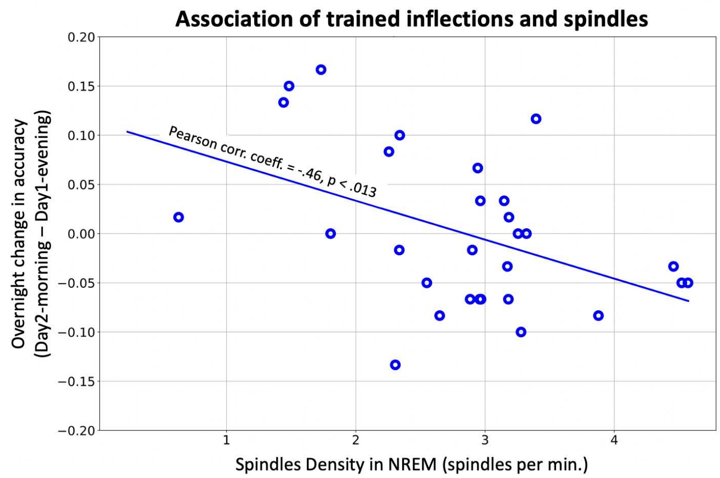

Figure 7. A negative association of the spindle density in NREM with the change in accuracy in the first interval: immediate to morning in the trained inflections task. Each dot denotes the data of one participant, the line is a group linear regression line.

Discussion

In this study we assessed the trajectories of consolidation of newly learned vocabulary and grammatical rules and their association with key aspects of sleep: SWS and sleep spindles. Across tasks, performance was above chance in all assessments, and higher exposure frequency benefited learning. However, we found differences in the temporal dynamics of consolidation between the different aspects of language learning. For vocabulary, accuracy significantly improved across the first night after learning; It then deteriorated between Day 3 and Day 9, but only for infrequent words. For trained plural inflections, accuracy improved between Day 2 and Day 3. In the generalisation test there was no change in accuracy across sessions. Overall, performance was highly correlated between the trained inflections and the generalisation tasks, but there were no correlations between the change in accuracy across intervals.

The associations with sleep metrics differed between the tasks. For vocabulary, SWS duration was positively associated with the overnight change (Day 1 to Day 2) in accuracy. Spindle density was positively associated with the change in vocabulary memory for infrequent words over the third interval (Day 3 to Day 9). For trained inflections, unexpectedly, spindle density was negatively associated with the change in accuracy over the first night post-training. There were no associations between sleep metrics and accuracy in generalisation.

Sleep and vocabulary acquisition

The overall increase in accuracy across the first night post-learning is consistent with previous studies showing a benefit of sleep to word learning (Tamminen et al., 2010; Walker et al., 2019), despite employing a different training and testing paradigm. This suggests that the benefit of sleep to vocabulary learning can be robustly measured across different paradigms as suggested by a recent review (Schimke, 2021). The positive association between overnight change and SWS duration (after controlling for total sleep time) is consistent with the suggestion that active consolidation processes during sleep benefited memory for vocabulary, as individuals with longer SWS exhibited larger accuracy benefits. Our results are consistent with those of a study by Tamminen and colleagues (2010) that showed a positive correlation between SWS duration and recognition speed of newly learned words, and more broadly with the association of declarative memory benefits with duration of SWS sleep (Gais & Born, 2004; Marshall & Born, 2007). As the current study did not include a wake control group, it remains possible that the observed improvement could have occurred without post-learning sleep; addressing this would require a direct test.

Mean group performance for infrequent words deteriorated across the third interval (Day 3 to Day 9), but on the individual level, this delayed decline was smaller in individuals with higher spindle density suggesting that endogenous replay of word-object pairs during the night after learning had long lasting effects on the retention of infrequent word-object pairs. Potentially, spindle activity reflected tagging for consolidation on subsequent nights (Cairney et al., 2018). Our data raise the possibility that tagging was specific to infrequent items (potentially due to weaker encoding; Denis et al., 2020; Drosopoulos et al., 2007; Schapiro et al., 2017), thus linking brain activity during the first post-learning night of sleep to performance for infrequent items eight days later. Another, non mutually exclusive, possibility is that the weaker encoding of infrequent words (due to the reduced exposure) makes this set of words more sensitive to forgetting, thereby allowing the benefits of replay during sleep to be revealed.

While the current study is the first to show this longer-term link between vocabulary acquisition and endogenous spindle activity, it is consistent with a previous report on the benefit of targeted memory reactivation following a serial reaction time task, to performance 10 days post-encoding, but not 24 hours post-encoding (Rakowska et al., 2021). Taken together, this suggests that active consolidation, via replay processes, protects word memories from being forgotten across longer time periods.

The specificity of the association of spindles with protection of infrequent words is also consistent with previous reports on prioritisation of weaker memories for offline consolidation, and specifically with association of spindles with the consolidation of weakly encoded memories (Denis et al., 2020; Drosopoulos et al., 2007; Petzka et al., 2021; Schmidt et al., 2006). For example, Denis and colleagues showed a positive association between fast sleep-spindle density and the consolidation of weakly encoded word-pairs over a six hour period that contained a nap (Denis et al., 2021). Our findings suggest that post-learning sleep contributes to the preferential strengthening of longer-term memory for words encountered less frequently. Taken together with the strong overall effect of exposure (i.e., high-frequency words were remembered significantly better than low-frequency ones), these results highlight the complementary roles of both exposure and sleep in natural language learning. For frequent words, extensive exposure may allow a substantial portion of learning to occur online. In contrast, for infrequent words, exposure alone may be insufficient to form robust representations, and thus, from computational and theoretical perspectives, these words could benefit more from offline consolidation. These offline benefits for infrequent words may be particularly important in human languages, where low-frequency words make up the vast majority of the vocabulary (Piantadosi, 2014). Our findings thus support the idea that sleep and exposure interact in a way that is optimally tuned to the structure of information in natural languages, thus supporting language acquisition.

Sleep and memory of trained plural inflections

At the group level, accuracy for trained inflections increased across the second interval - which did not include the first night of sleep after learning. One might therefore expect spindle density to be positively correlated with the change in accuracy across this interval; however, our data did not support this hypothesis.

We also found an unexpected negative association between spindles and overnight change in accuracy for trained inflections. This finding is in line with a study by Lustenberger and colleagues (2012) who found that fast spindle activity had a negative correlation with overnight change in performance in a word pairs task. In their study, spindle activity also positively correlated with immediate post-learning performance, suggesting that participants who generally have more spindles, are better learners who achieve more of their maximal capacity during online encoding in the evening, and thus show less overnight improvement. Our data provide only partial support for this suggestion: The association of pre-sleep performance and overnight change was indeed negative. However, there was no positive association between spindle density and pre-sleep performance, and thus we cannot conclude that participants with more spindles also showed better encoding.

Finally, it is important to note that our study included PSG during a single night, and therefore we cannot differentiate between individual baseline "trait" level spindle density and changes in spindle density due to learning ("state"; Gais et al., 2002). These two components have been shown to exhibit distinct patterns of correlations with behavioural measures of learning (Lustenberger et al., 2015; Schabus et al., 2004, 2008; Schmidt et al., 2006), suggesting a functional distinction. Thus, it might be the case that participants who had lower spindle density as measured in our study were in fact participants whose baseline spindle levels are low, but for whom the change in density that is associated with a learning experience was high (whereas participants for whom we measured higher spindle density were participants whose baseline spindle levels are high, and the learning-induced change was low). Taking this into account, it is theoretically plausible that participants with greater learning-related changes in spindle density were those who showed higher accuracy gains, in line with previous findings on word-pair learning consolidation (Schmidt et al., 2006). Further research spanning multiple nights is needed to allow individual measurement of post-learning spindle changes relative to a baseline, and the association of these two metrics with change in accuracy.

Sleep and extraction of linguistic regularities

For linguistic regularity extraction, as measured by the generalisation test, mean group performance was above chance across time points, with no significant changes between them. Examining individual differences we did not find a correlation between overnight change in accuracy and SWS duration or sleep spindles.

The lack of significant improvement across the different timepoints is consistent with a previous study that examined delayed generalisation (Mirković et al., 2019), but seems to be at odds with a previous study that used the same training procedure as we used here and found small but significant improvement across 24 hours (Ben-Zion et al., 2022). However, in the study by Ben-Zion and colleagues, no group difference was found between a group that slept shortly after training (PM training) and a group who did not (AM training), at 12 and 24 hours measurements, thus showing no evidence that the extraction of regularities depends on sleep. Similarly to the experimental paradigm in the current study, the generalisation in these studies involved production. However, the type of access involved in production does not seem to be the factor masking sleep-related benefits for rule learning. Mirković & Gaskell (2016) examined the extraction of grammatical regularities using a paradigm in which the generalisation test did not involve production, and reported that a short nap did not enhance the extraction of language regularities more than a period of wakefulness. Furthermore, Tamminen (2020) showed that learning of a new writing system, including regularity extraction, can withstand sleep deprivation.

However, others have found that sleep benefits the acquisition of word order rules (Cross et al., 2024), artificial grammar acquisition by infants (Gómez et al., 2006) and adults (Nieuwenhuis et al., 2013), learning of phonotactic constraints in speech production (Gaskell et al., 2014), and generalisation to novel input in synthetic speech perception (Fenn et al., 2003). Furthermore, some associations between sleep metrics and regularity extraction in language have been documented (Batterink et al., 2014; Batterink & Paller, 2017). For example, in Batterink et al. (2014) participants acquired an implicit rule for using novel articles, and there was no significant change in group mean performance after the nap as compared to before the nap, in alignment with the results of the current study. However, participants who had more *slow-wave sleep duration x REM duration* during the nap, showed a greater increase in sensitivity to the hidden linguistic rule between the two experimental sessions. Given this evidence, the conclusions on grammar acquisition are less clear-cut compared to vocabulary acquisition; We will return to this question in the following section.

<u>Integrating results across tasks and their associations with sleep</u>

While performance in the generalisation task has to rely on knowledge of the regularities or constraints that underlie plural inflections, performance in the trained inflections task may be supported by two non-mutually exclusive factors: memory of the forms for specific pairs of singular and plurals, and knowledge of the plural inflectional regularities. Each of these factors, theoretically, can support a correct response to all trials. The significant correlations across all timepoints between performance on trained inflections and generalisation (also reported by Ben-Zion and colleagues who used a similar paradigm; Ben Zion et al., 2019; Ben-Zion et al., 2022) fit with the proposal that knowledge of the plural regularities/constraints is the dominant mechanism employed in the trained inflections task. However, it could also be the case that individuals who are better in learning the set of the plural forms that they were exposed to during training, are also better in learning the rules that underlie these plural forms. This means that performance in these two tasks does not necessarily rely on a fully shared mechanism. For instance, knowledge of word structure facilitates the acquisition of novel words (Anglin et al., 1993; Carlisle, 2000; Mahony et al., 2000), nonetheless, specific lexical knowledge and morphological rule knowledge remain distinct constructs.

The consolidation dynamics for trained inflections followed a different pattern as compared to the generalisation task. First, performance for generalisation did not significantly change across timepoints, whereas performance for trained inflections improved across the second day. Second, there were no correlations between the change in performance in the two tasks across the examined time intervals (replicating previous findings using a similar paradigm; Ben Zion et al., 2019; Ben-Zion et al., 2022). Furthermore, generalisation was not found to be associated with the examined sleep metrics, whereas the change in performance for trained inflections across the night post-learning was negatively associated with sleep-spindle density. Thus, even if retrieval in the two tasks partially relies on overlapping representations, the formation of these representations seems to be associated with distinct neural mechanisms.

The significant correlation between vocabulary and generalisation on Days 2 and 3 may be linked to the fact that neither was an explicit target of learning: In the training phase, participants were only required to produce the plural form of the presented words. It is also worth noting that although it was not statistically significant, all within-session correlations showed a positive trend.

Our findings are consistent with the idea that vocabulary learning, which is associated with the episodic, hippocampus-dependent system, is stabilised by sleep. In contrast, the extraction of regularities may depend less on the hippocampus and instead rely more on frontostriatal skill-learning circuitry (Gaskell, 2024; Ullman, 2016), and is therefore supported by sleep to a lesser extent. Indeed, in a neuroimaging study by Nevat and colleagues (2017) that used a very similar paradigm to the one used in the current study, the frontostriatal network was activated during inflection of trained items with no involvement of medial temporal structures. Our findings that vocabulary knowledge, but not trained inflection or generalisation, improved during the first post-learning night align with the idea of greater initial dependency of novel vocabulary on the hippocampus and, consequently, a greater benefit from sleep.

The training procedure used in the current study comprised both arbitrary language aspects (i.e., vocabulary - the semantics of the stem), and systematic aspects (i.e., the implicit morpho-phonological regularity). It has been previously suggested that when both aspects are learned simultaneously, as part of the same procedure, systematic aspects might not show sleep-related benefits due to a prioritisation of consolidation of the arbitrary components during post-learning sleep (Mirković & Gaskell, 2016; Sweegers & Talamini, 2014). That is because arbitrary aspects are thought to be most dependent on the hippocampus during initial encoding and so they are being prioritised during sleep initially, while systematic aspects being prioritised later on (McClelland et al., 1995; Mirković & Gaskell, 2016; Stickgold & Walker, 2013). Indeed, in a study that found a clear sleep vs. wake benefit to syntactic rule acquisition (when participants were aware of the rule before sleep), no arbitrary aspects of language were part of the training as the sentences consisted of existing English words (Kim & Fenn, 2020). This could potentially explain our current finding of a lack of improvement in the trained inflections and the generalisation tasks over the first night of training, in contrast to the improvement in vocabulary.

We found a negative association of spindle density with the change in accuracy in the trained inflections task. A possible interpretation is that consolidation resources, quantified in this study by SWS duration and spindle density, were allocated to label-object pairings (i.e., vocabulary task) more than to stem-plural form pairings (i.e., trained inflections task). In line with this suggestion, Antony and colleagues (2018) showed that cuing during sleep had a detrimental effect on memory of picture-location pairs when these were learned in a competitive condition. In our data, this suggestion is supported by the positive correlation between SWS and overnight change in memory for vocabulary, taken together with the positive correlation between spindles and protection of infrequent words over the delayed period vs. the negative correlation between spindles and overnight change in trained inflections. Importantly, a single-night PSG does not allow separating the “trait” spindle activity of an individual from the change following learning (i.e., “state”). Thus, it could be the case that the positive correlation between spindles and vocabulary stems from a correlation with the changing component in overall spindle activity, whereas the negative correlation with trained inflections stems from a correlation with the baseline component. This suggestion is further supported by the weak correlations between performance for vocabulary and trained inflections tasks.

Finally, although both SWS and spindle density are part of memory replay mechanisms, the correlations between accuracy in the different tasks with SWS duration and sleep spindles, varied in our data (see also Tamminen et al., 2010). This suggests a temporally distinct role for SWS vs. spindles: SWS might be related to immediate consolidation of vocabulary, while spindles might mediate longer-term consolidation and protection against forgetting of low-frequency items.

Studying these interactions is especially important given that the co-occurrence of sleep spindles with slow oscillations (i.e., coupled spindles) has been shown to benefit memory (Denis & Cairney, 2023; Klinzing et al., 2019; Staresina, 2024) and may be specifically associated with consolidation of weakly encoded memories (Denis et al., 2021). This, taken together with our results, highlights the need to examine SWS, and coupled and uncoupled sleep spindles within the same study in order to develop a more detailed understanding of their distinct roles in memory consolidation.

Although these findings do not directly inform language teaching practices or interventions, they underscore the importance of considering both encoding and consolidation when evaluating teaching or intervention outcomes, in typical populations and in individuals with learning difficulties. In particular, they point to the importance of delayed assessment, especially for vocabulary and low-frequency words.

Limitations

The findings of this study are subject to a number of limitations. The study employed a small artificial language learned under laboratory conditions, which may raise some questions about the relevance of the findings to natural language learning, although several aspects of the design support its broader applicability. First, mechanisms of encoding and consolidation, particularly in studies examining individual differences, are presumably activated also in a laboratory setting. Second, the artificial language incorporated properties of natural language, such as semantics (each word had a meaning), a linguistically plausible plural suffix rule system, and irregularities within the rule system. Nonetheless, factors such as the learning context and the ecological relevance of the material are inherent limitations in this type of study.

PSG data were collected over only one night of sleep, thus not allowing the separation of baseline neural activity during sleep from specific post-learning change. We also did not include a wake control group due to a number of factors: (1) The overall complexity of the study, (2) The availability of prior evidence on the specific contribution of sleep from a wake vs. sleep study that used a very similar paradigm (Ben-Zion et al., 2022), (3) Study design: We assessed performance on the behavioural tasks in relation to specific sleep metrics, thereby linking behavioural outcomes directly to sleep physiology.

Another limitation of the study is that the generalisation test was not administered on Day 9 due to technical reasons. While the available data are consistent with the view that memory for rules does not benefit from sleep, they are also compatible with the suggestion that the benefit across time is small (Ben-Zion et al., 2022) but consistent, and may accumulate over longer time periods. We aim to address the question of longer-term rule learning and its dependency on sleep and additional exposure in future studies.

For vocabulary, we report a significant three-way interaction with a moderate effect between word frequency, testing time and spindle density. However, the binary nature of the task taken together with our sample size and the complexity of the model, may limit the stability and generalisability of this effect.

Finally, we assessed associations with two well-established sleep metrics: SWS duration and sleep spindle density. We limited ourselves to these measures in order to avoid a proliferation of tests and reduce the risk of false positives due to multiple comparisons (Ranganathan et al., 2016). However, additional measures such as coupled spindles, slow-wave activity, and slow oscillation density could potentially contribute further to our understanding of the mechanisms underlying active memory consolidation.

<u>Main Contributions</u>

In this study, we investigated the temporal dynamics of language learning across nine days and examined its relationship to two key memory-related sleep characteristics: sleep spindles and slow-wave sleep (SWS) duration. On the group level, memory for vocabulary improved over the first night post-learning, memory for trained plural forms improved over the second day post-learning, and there was no change in generalisation up to three days post-training. On the individual level, sleep metrics were associated more with vocabulary learning than with rule learning: SWS duration was positively correlated with vocabulary consolidation across the first night post-learning, and sleep spindles showed a potential protective role for longer-term retention of learnt infrequent words. The latter is especially thought provoking as the vast majority of words in human languages are infrequent, rendering a mechanism like this highly beneficial to everyday language learning. We found a negative association of spindles with changes in plural inflections and no associations of sleep metrics with changes in generalisation, warranting future investigation.

The study design offered several unique strengths: (1) It addressed both arbitrary and systematic aspects of language learning, (2) It assessed two core sleep-related learning metrics within the same group of participants, enabling direct comparisons between them, (3) It employed an extended timescale of testing, and (4) It is the first study to directly link post-learning neural activity during sleep with longer-term learning outcomes. This work highlights the multifaceted role of sleep in language learning and emphasises the importance of investigating how post-learning slow-wave sleep (SWS) and sleep spindles contribute to consolidation processes across extended time periods and varying types of linguistic knowledge.

<u>Declaration of generative AI and AI-assisted technologies in the writing process</u>

During the preparation of this work the authors used ChatGPT in order to improve phrasing and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Bibliography

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary Development: A Morphological Analysis. *Monographs of the Society for Research in Child Development*, *58*(10).

Antony, J. W., Cheng, L. Y., Brooks, P. P., Paller, K. A., & Norman, K. A. (2018). Competitive learning modulates memory consolidation during sleep. *Neurobiology of Learning and Memory*, *155*, 216–230. https://doi.org/10.1016/j.nlm.2018.08.007

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, *65*, 169–179. https://doi.org/10.1016/j.neuropsychologia.2014.10.024

Batterink, L. J., & Paller, K. A. (2017). Sleep-Based Memory Processing Facilitates Grammatical Generalization: Evidence from Targeted Memory Reactivation. *Brain Language*, *167*, 83–93. https://doi.org/10.1016/j.bandl.2015.09.003.Sleep-Based

Becker, S. P. (2020). ADHD and sleep: Recent advances and future directions. *Current Opinion in Psychology*, *34*, 50–56. https://doi.org/10.1016/j.copsyc.2019.09.006

Ben Zion, D., Nevat, M., Prior, A., & Bitan, T. (2019). Prior knowledge predicts early consolidation in second language learning. *Frontiers in Psychology*, *10*(OCT), 1–15. https://doi.org/10.3389/fpsyg.2019.02312

Ben-Zion, D., Gabitov, E., Bitan, T., & Prior, A. (2023). Impaired extraction and consolidation of morphological regularities in developmental dyslexia: A domain general deficit? *Neuropsychologia*, *188*, 108652. https://doi.org/10.1016/j.neuropsychologia.2023.108652

Ben-Zion, D., Gabitov, E., Prior, A., & Bitan, T. (2022). Effects of Sleep on Language and

Motor Consolidation: Evidence of Domain General and Specific Mechanisms. *Neurobiology of Language*. https://doi.org/10.1162/nol_a

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant’s Age. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01116

Cairney, S. A., Guttesen, A. Á. V., El Marj, N., & Staresina, B. P. (2018). Memory Consolidation Is Linked to Spindle-Mediated Information Processing during Sleep. *Current Biology*, *28*(6), 948-954.e4. https://doi.org/10.1016/j.cub.2018.01.087

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal*, *12*, 169–190.

Clay, F., Bowers, J. S., Davis, C. J., & Hanley, D. A. (2007). Teaching adults new words: The role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 970–976. https://doi.org/10.1037/0278-7393.33.5.970

Clemens, Z., Fabó, D., & Halász, P. (2005). Overnight verbal memory retention correlates with the number of sleep spindles. *Neuroscience*, *132*(2), 529–535. https://doi.org/10.1016/j.neuroscience.2005.01.011

Cordi, M. J., & Rasch, B. (2021). How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology*, *67*, 1–7. https://doi.org/10.1016/j.conb.2020.06.002

Cross, Z. R., Helfrich, R. F., Corcoran, A. W., Dede, A. J. O., Kohler, M. J., Coussens, S. W., Zou-Williams, L., Schlesewsky, M., Gaskell, M. G., Knight, R. T., & Bornkessel-Schlesewsky, I. (2024). Slow oscillation-spindle coupling predicts sequence-based language learning. *The Journal of Neuroscience*, e2193232024. https://doi.org/10.1523/JNEUROSCI.2193-23.2024

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, *21*(4), 803–820. https://doi.org/10.1162/jocn.2009.21059

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

Denis, D., & Cairney, S. A. (2023). Neural reactivation during human sleep. *Emerging Topics in Life Sciences*, *7*(5), 487–498. https://doi.org/10.1042/ETLS20230109

Denis, D., Mylonas, D., Poskanzer, C., Bursal, V., Payne, J. D., & Stickgold, R. (2021). Sleep Spindles Preferentially Consolidate Weakly Encoded Memories. *The Journal of Neuroscience*, *41*(18), 4088–4099. https://doi.org/10.1523/JNEUROSCI.0818-20.2021

Denis, D., Schapiro, A. C., Poskanzer, C., Bursal, V., Charon, L., Morgan, A., & Stickgold, R. (2020). The roles of item exposure and visualization success in the consolidation of memories across wake and sleep. *Learning & Memory*, *27*, 451–456.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*(2), 114–126. https://doi.org/10.1038/nrn2762

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*(5), 309–321. https://doi.org/10.1016/j.smrv.2008.08.002

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), 169–183. https://doi.org/10.1037/0096-3445.136.2.169

Dumay, N., & Gaskell, M. G. (2007). Sleep-Associated Changes in the Mental

Representation of Spoken Words. *PSYCHOLOGICAL SCIENCE Research*, *18*(1), 35–39.

Durrant, S. J., Cairney, S. A., & Lewis, P. A. (2013). Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum. *Cerebral Cortex*, *23*(10), 2467–2478. https://doi.org/10.1093/cercor/bhs244

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia*, *49*(5), 1322–1331. https://doi.org/10.1016/j.neuropsychologia.2011.02.015

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*(18), 7723–7728. https://doi.org/10.1073/pnas.0700094104

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*(6958), 614–616. https://doi.org/10.1038/nature01951

Gais, S., & Born, J. (2004). Declarative memory consolidation: Mechanisms acting during human sleep. *Learning & Memory*, *11*(6), 679–685. https://doi.org/10.1101/lm.80504

Gais, S., Mölle, M., Helms, K., & Born, J. (2002). Learning-dependent increases in sleep spindle density. *Journal of Neuroscience*, *22*(15), 6830–6834. https://doi.org/10.1523/jneurosci.22-15-06830.2002

Gaskell, M. G. (2024). EPS mid-career prize: An integrated framework for the learning, recognition and interpretation of words. *Quarterly Journal of Experimental Psychology*, *77*(12), 2365–2384. https://doi.org/10.1177/17470218241284289

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

Gaskell, M. G., Warker, J., Lindsay, S., Frost, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep Underpins the Plasticity of Language Production. *Psychological*

Science, *25*(7), 1457–1465. https://doi.org/10.1177/0956797614535937

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps Promote Abstraction in Language-Learning Infants. *Psychological Science*, *17*(8), 670–674. https://doi.org/10.1111/j.1467-9280.2006.01764.x

Gramfort, A. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Horne, J. A., & Ostberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*(2), 97–110.

Inc, T. M. (2022). *MATLAB version: 9.13.0 (R2022b)* [Computer software]. The MathWorks Inc. https://www.mathworks.com

James, E., Gaskell, M. G., & Henderson, L. M. (2019). *Offline consolidation supersedes prior knowledge benefits in children ' s (but not adults ') word learning*. *March 2018*, 1–19. https://doi.org/10.1111/desc.12776

Kim, K. M., & Fenn, K. M. (2020). SLEEP-DEPENDENT CONSOLIDATION OF SECOND LANGUAGE GRAMMAR KNOWLEDGE: THE ROLE OF AWARENESS. *Studies in Second Language Acquisition*, *42*(5), 1107–1120. https://doi.org/10.1017/S0272263120000200

Klinzing, J. G., Niethard, N., & Born, J. (2019). Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, *22*(10), 1598–1610. https://doi.org/10.1038/s41593-019-0467-3

Konofal, E., Lecendreux, M., & Cortese, S. (2010). Sleep and ADHD. *Sleep Medicine*, *11*(7), 652–658. https://doi.org/10.1016/j.sleep.2010.02.012

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do

Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Kumral, D., Matzerath, A., Leonhart, R., & Schönauer, M. (2023). Spindle-dependent memory consolidation in healthy adults: A meta-analysis. *Neuropsychologia*, *189*, 108661. https://doi.org/10.1016/j.neuropsychologia.2023.108661

Leach, S., Krugliakova, E., Sousouri, G., Snipes, S., Skorucak, J., Schühle, S., Müller, M., Ferster, M. L., Da Poian, G., Karlen, W., & Huber, R. (2024). Acoustically evoked K-complexes together with sleep spindles boost verbal declarative memory consolidation in healthy adults. *Scientific Reports*, *14*(1), 19184. https://doi.org/10.1038/s41598-024-67701-7

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, *15*(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Lustenberger, C., Maric, A., Dürr, R., Achermann, P., & Huber, R. (2012). Triangular Relationship between Sleep Spindle Activity, General Cognitive Ability and the Efficiency of Declarative Learning. *PLoS ONE*, *7*(11), e49561. https://doi.org/10.1371/journal.pone.0049561

Lustenberger, C., Wehrle, F., Tüshaus, L., Achermann, P., & Huber, R. (2015). The Multidimensional Aspects of Sleep Spindles and Their Relationship to Word-Pair Memory Consolidation. *Sleep*, *38*(7), 1093–1103. https://doi.org/10.5665/sleep.4820

Mahony, D., Singson, M., & Mann, V. (2000). Reading Ability and Sensitivity to Morphological Relations. *Reading and Writing: An Interdisciplinary Journal*, *12*, 191–218. https://doi.org/10.1023/A:1008136012492

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences*, *11*(10), 442–450.

https://doi.org/10.1016/j.tics.2007.09.001
McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.
Mirković, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PLoS ONE*, *11*(4), 1–26. https://doi.org/10.1371/journal.pone.0152489
Mirković, J., Vinals, L., & Gaskell, M. G. (2019). The role of complementary learning systems in learning and consolidation in a quasi-regular domain. *Cortex*, *116*, 228–249. https://doi.org/10.1016/j.cortex.2018.07.015
Mölle, M., Bergmann, T. O., Marshall, L., & Born, J. (2011). Fast and slow spindles during the sleep slow oscillation: Disparate coalescence and engagement in memory processing. *Sleep*, *34*(10), 1411–1421. https://doi.org/10.5665/SLEEP.1290
Natale, V., Fabbri, M., Tonetti, L., & Martoni, M. (2014). Psychometric goodness of the Mini Sleep Questionnaire. *Psychiatry and Clinical Neurosciences*, *68*(7), 568–573. https://doi.org/10.1111/pcn.12161
Nevat, M., Ullman, M. T., Eviatar, Z., & Bitan, T. (2017). The neural bases of the learning and generalization of morphological inflection. *Neuropsychologia*, *98*(August 2016), 139–155. https://doi.org/10.1016/j.neuropsychologia.2016.08.026
Nevat, M., Ullman, M. T., Eviatar, Z., & Bitan, T. (2018). The role of distributional factors in learning and generalising affixal plural inflection: An artificial language study. *Language, Cognition and Neuroscience*, *33*(9), 1184–1204. https://doi.org/10.1080/23273798.2018.1465187
Ng, T., Noh, E., & Spencer, R. M. C. (2024). *Does slow oscillation-spindle coupling contribute to sleep-dependent memory consolidation? A Bayesian meta-analysis*.

https://doi.org/10.1101/2024.08.28.610060

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep Promotes the Extraction of Grammatical Rules. *PLOS ONE*, *8*(6), 1–10. https://doi.org/10.1371/journal.pone.0065046

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary Learning Systems. *Cognitive Science*, *38*(6), 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Palma, P., & Titone, D. (2021). Something old, something new: A review of the literature on sleep-related lexicalization of novel words in adults. *Psychonomic Bulletin & Review*, *28*(1), 96–121. https://doi.org/10.3758/s13423-020-01809-5

Petzka, M., Charest, I., Balanos, G. M., & Staresina, B. P. (2021). Does sleep-dependent consolidation favour weak memories? *Cortex*, *134*, 65–75. https://doi.org/10.1016/j.cortex.2020.10.005

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rakowska, M., Abdellahi, M. E. A., Bagrowska, P., Navarrete, M., & Lewis, P. A. (2021). Long term effects of cueing procedural memory reactivation during NREM sleep. *NeuroImage*, *244*, 118573. https://doi.org/10.1016/j.neuroimage.2021.118573

Ranganathan, P., Pramesh, C., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in Clinical Research*, *7*(2), 106. https://doi.org/10.4103/2229-3485.179436

Rucklidge, J. J., & Tannock, R. (2002). Neuropsychological profiles of adolescents with ADHD: Effects of reading difficulties and gender. *Journal of Child Psychology and*

Psychiatry, *43*(8), 988–1003. https://doi.org/10.1111/1469-7610.00227

Schabus, M., Gruber, G., Parapatics, S., Sauter, C., Klösch, G., Anderer, P., Klimesch, W., Saletu, B., & Zeitlhofer, J. (2004). Sleep spindles and their significance for declarative memory consolidation. *Sleep*, *27*(8), 1479–1485. https://doi.org/10.1093/sleep/27.7.1479

Schabus, M., Hoedlmoser, K., Pecherstorfer, T., Anderer, P., Gruber, G., Parapatics, S., Sauter, C., Kloesch, G., Klimesch, W., Saletu, B., & Zeitlhofer, J. (2008). Interindividual sleep spindle differences and their relation to learning-related enhancements. *Brain Research*, *1191*, 127–135. https://doi.org/10.1016/j.brainres.2007.10.106

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep Benefits Memory for Semantic Category Structure While Preserving Exemplar-Specific Information. *Scientific Reports*, *7*(1), 14869. https://doi.org/10.1038/s41598-017-12884-5

Schimke, E. A. E., Angwin, A. J., Cheng, B. B. Y., & Copland, D. A. (2021). The effect of sleep on novel word learning in healthy adults: A systematic review and meta-analysis. *Psychonomic Bulletin & Review*, *28*(6), 1811–1838. https://doi.org/10.3758/s13423-021-01980-3

Schmidt, C., Peigneux, P., Muto, V., Schenkel, M., Knoblauch, V., Münch, M., De Quervain, D. J.-F., Wirz-Justice, A., & Cajochen, C. (2006). Encoding Difficulty Promotes Postlearning Changes in Sleep Spindle Activity during Napping. *The Journal of Neuroscience*, *26*(35), 8976–8982. https://doi.org/10.1523/JNEUROSCI.2464-06.2006

Staresina, B. P. (2024). Coupled sleep rhythms for memory consolidation. *Trends in Cognitive Sciences*, *28*(4), 339–351. https://doi.org/10.1016/j.tics.2024.02.002

Stickgold, R., & Walker, M. (2005). Memory consolidation and reconsolidation: What is the role of sleep? *Trends in Neurosciences*, *28*(8), 408–415. https://doi.org/10.1016/j.tins.2005.06.004

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience*, *16*(2), 139–145. https://doi.org/10.1038/nn.3303.Sleep-dependent

Sweegers, C. C. G., & Talamini, L. M. (2014). Generalization from episodic memories across time: A route for semantic knowledge acquisition. *Cortex*, *59*, 49–61. https://doi.org/10.1016/j.cortex.2014.07.006

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112. https://doi.org/10.1016/j.cognition.2012.06.014

Tamminen, J., & Gaskell, M. G. (2008). Short Article: Newly Learned Spoken Words Show Long-Term Lexical Competition Effects. *Quarterly Journal of Experimental Psychology*, *61*(3), 361–371. https://doi.org/10.1080/17470210701634545

Tamminen, J., & Gaskell, M. G. (2012). Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology (2006)*, *October*, 37–41. https://doi.org/10.1080/17470218.2012.724694

Tamminen, J., Lambon Ralph, M. A., & Lewis, P. A. (2013). The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *Journal of Neuroscience*, *33*(39), 15376–15381. https://doi.org/10.1523/JNEUROSCI.5093-12.2013

Tamminen, J., Newbury, C. R., Crowley, R., Vinals, L., Cevoli, B., & Rastle, K. (2020). Generalisation in language learning can withstand total sleep deprivation. *Neurobiology of Learning and Memory*, *173*(July), 107274. https://doi.org/10.1016/j.nlm.2020.107274

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. *The Journal of Neuroscience*, *30*(43), 14356–14360. https://doi.org/10.1523/JNEUROSCI.3028-10.2010

Tham, E. K. H., Lindsay, S., & Gaskell, M. G. (2015). Markers of automaticity in sleep-associated consolidation of novel words. *Neuropsychologia*, *71*, 146–157. https://doi.org/10.1016/j.neuropsychologia.2015.03.025

Ullman, M. T. (2016). The Declarative/Procedural Model. In *Neurobiology of Language* (pp. 953–968). Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Vallat, R., & Walker, M. P. (2021). An open-source, high-performance tool for automated sleep staging. *eLife*, *10*, e70092. https://doi.org/10.7554/eLife.70092

Van Der Ven, F., Takashima, A., Segers, E., & Verhoeven, L. (2015). Learning Word Meanings: Overnight Integration and Study Modality Effects. *PLOS ONE*, *10*(5), e0124926. https://doi.org/10.1371/journal.pone.0124926

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*(6972), 352–355. https://doi.org/10.1038/nature02223

Walker, S., Henderson, L. M., Fletcher, F. E., Knowland, V. C. P., Cairney, S. A., & Gaskell, M. G. (2019). Learning to live with interfering neighbours: The influence of time of learning and level of encoding on word learning. *Royal Society Open Science*, *6*(4), 181842. https://doi.org/10.1098/rsos.181842

Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A., & Gaskell, M. G. (2017). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and Language*, *167*, 13–27. https://doi.org/10.1016/j.bandl.2016.07.010

Zohar, A. H., & Konfortes, H. (2010). Diagnosing ADHD in Israeli adults: The psychometric properties of the adult ADHD self report scale (ASRS) in hebrew. *Israel Journal of Psychiatry and Related Sciences*, *47*(4), 308–313.

Supplementary

Table S1A. Vocabulary learning: Accuracy as predicted by word frequency and time of assessment.

Full results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1 + word_freq * time_point + (1 + word_freq | participant).* time_point denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9 after. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

$^{*}\ p < .05$, $^{**}\ p < 10^{-4}$

Predictor	**β**	**SE**	**z value**	**p**
word_freq2-1	-0.29	0.08	-3.88	10^{-4} **
time_point2-1	0.20	0.08	2.55	0.011 *
time_point3-2	0.07	0.08	0.87	0.383
time_point4-3	-0.17	0.08	-2.13	0.033 *
word_freq2-1:time_point2-1	0.18	0.16	1.13	0.259
word_freq2-1:time_point3-2	0.04	0.16	0.26	0.796
word_freq2-1:time_point4-3	-0.35	0.16	-2.20	0.028 *

Table S1B. Vocabulary learning: Response times as predicted by word frequency and time of assessment.

Full results of the linear mixed-regression model that Buildmer converged to: *logRT ~ 1 + frequency x time point + (1 + time point | participant).* time_point denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *frequency*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < .05$, ** $p < 10^{-4}$

Predictor	**Estimate**	**SE**	**t value**	**p**
frequency2-1	0.02	0.01	2.13	0.033 *
time_point2-1	-0.04	0.02	-1.67	0.106
time_point3-2	-0.07	0.01	-4.45	10^{-4} **
time_point4-3	-0.04	0.02	-2.62	0.014 *
frequency2-1:time_point2-1	-0.02	0.02	-0.87	0.386
frequency2-1:time_point3-2	0.02	0.02	0.82	0.414
frequency2-1:time_point4-3	-0.02	0.02	-1.16	0.245

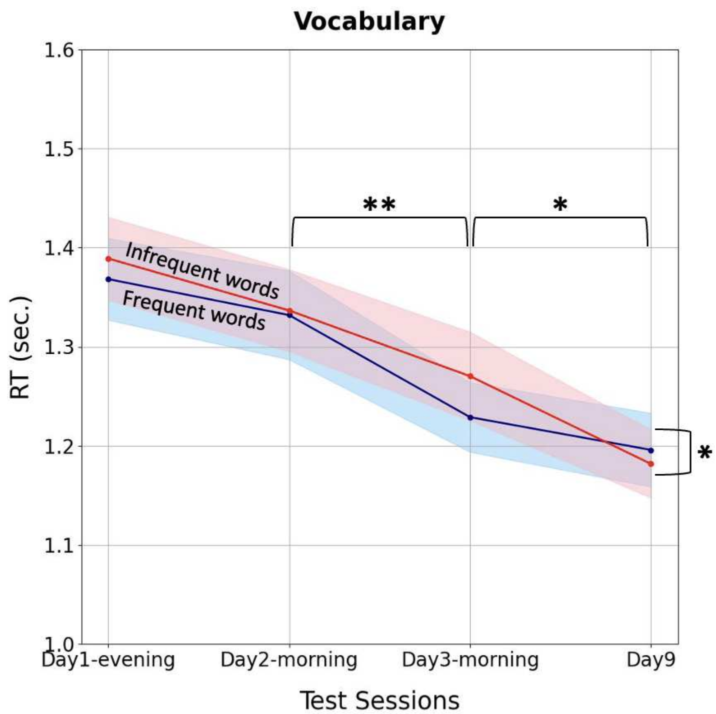

Figure S1. Average response times for the vocabulary task in the four time points; shaded areas denote standard errors.

Table S1C. Trained inflections: Accuracy as predicted by word frequency and time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1 + frequency * time_of_assessment + (1 + frequency | participant).* Time_point denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

$* \ p < .05$, $** \ p < 10^{-4}$

Predictor	**β**	**SE**	**z value**	**p**
word_freq2-1	-0.34	0.08	-4.27	10^{-4} **
time_point2-1	-0.01	0.08	-0.11	0.915
time_point3-2	0.21	0.09	2.46	0.014 *
time_point4-3	0.04	0.09	0.41	0.683
word_freq2-1:time_point2-1	-0.14	0.17	-0.84	0.401
word_freq2-1:time_point3-2	0.12	0.17	0.72	0.471
word_freq2-1:time_point4-3	-0.19	0.18	-1.10	0.273

Table S1D. Trained inflections: Response times as predicted by word frequency and time of assessment.

Full results of the linear mixed-regression model that Buildmer converged to: *logRT ~ 1 + word frequency x time point + (1 + time point | participant).* time_point denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < 10^{-8}$

Predictor	**Estimate**	**SE**	**t value**	**p**
word_freq2-1	0.14	0.02	5.73	10^{-8} *
time_point2-1	0.01	0.06	0.13	0.901
time_point3-2	-0.06	0.07	-0.96	0.347
time_point4-3	-0.06	0.04	-1.46	0.154
word_freq2-1:time_point2-1	-0.01	0.07	-0.16	0.874
word_freq2-1:time_point3-2	-0.05	0.07	-0.64	0.522
word_freq2-1:time_point4-3	-0.03	0.07	-0.36	0.719

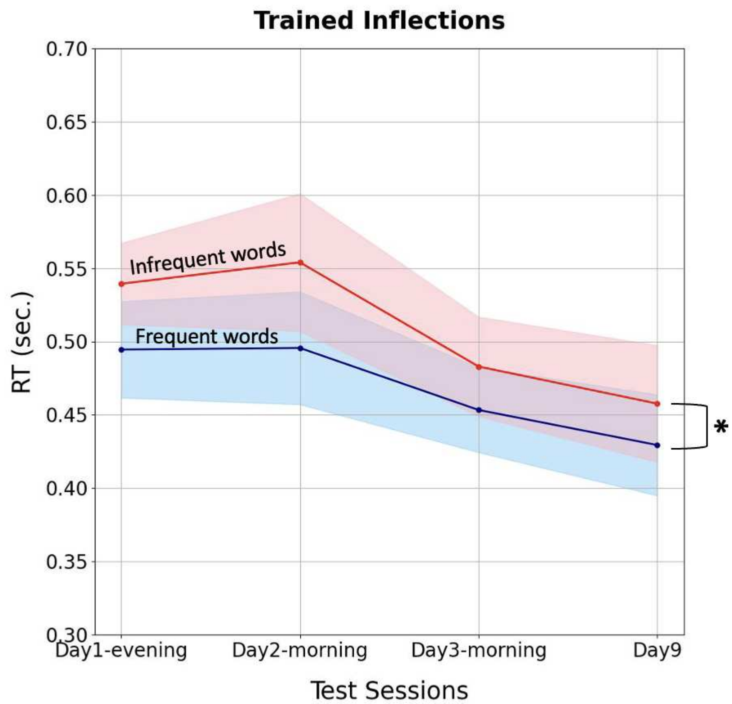

Figure S2. Average response times for the trained inflections task in the four time points; shaded areas denote standard errors.

Table S1E. Generalisation: Accuracy of production as predicted by time of assessment. Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1* + time_point + *(1 | participant)*. time_point denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*.

Predictor	**β**	**SE**	**z value**	**p**
time_point2-1	0.14172	0.10429	1.359	0.174
time_point3-2	0.06587	0.10462	0.63	0.529

Table S2A. Total sleep time and SWS duration (in minutes) as predictors of performance in the vocabulary task, in addition to word frequency and time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1 + word_freq * time_of_assessment * total_sleep_length + word_freq * time_of_assessment * SWS_length + (1 + word_freq | participant)*. *time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < .05$, ** $p < 10^{-4}$

Predictor	**β**	**SE**	**z value**	**p**
word_freq2-1	-0.29	0.07	-3.93	10^{-4} **
time_point2-1	0.19	0.08	2.43	0.015 *
time_point3-2	0.07	0.08	0.91	0.361
time_point4-3	-0.17	0.08	-2.16	0.031 *
total_sleep_length_cent	-0.04	0.10	-0.43	0.669
SWS_min_cent	-0.07	0.10	-0.76	0.449
word_freq2-1:time_point2-1	0.18	0.16	1.11	0.266
word_freq2-1:time_point3-2	0.03	0.16	0.21	0.833
word_freq2-1:time_point4-3	-0.35	0.16	-2.16	0.031 *
word_freq2-1:total_sleep_length_cent	-0.03	0.07	-0.43	0.670
time_point2-1:total_sleep_length_cent	-0.01	0.08	-0.11	0.916
time_point3-2:total_sleep_length_cent	-0.06	0.08	-0.72	0.473
time_point4-3:total_sleep_length_cent	0.04	0.08	0.52	0.606
word_freq2-1:SWS_min_cent	0.06	0.08	0.78	0.438
time_point2-1:SWS_min_cent	0.20	0.08	2.39	0.017 *

time_point3-2:SWS_min_cent	0.00	0.08	-0.01	0.992
time_point4-3:SWS_min_cent	-0.05	0.08	-0.58	0.561
word_freq2-1:time_point2-1:total_sleep_length_cent	0.02	0.16	0.15	0.879
word_freq2-1:time_point3-2:total_sleep_length_cent	0.20	0.16	1.24	0.215
word_freq2-1:time_point4-3:total_sleep_length_cent	-0.18	0.16	-1.13	0.258
word_freq2-1:time_point2-1:SWS_min_cent	0.09	0.16	0.58	0.564
word_freq2-1:time_point3-2:SWS_min_cent	0.07	0.17	0.40	0.692
word_freq2-1:time_point4-3:SWS_min_cent	-0.07	0.17	-0.45	0.654

Table S2B. Total sleep time and duration of SWS as predictors of performance in the trained inflections task, in addition to word frequency and time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1 + word_freq * time_of_assessment * total_sleep_duration + word_freq * time_of_assessment * SWS_duration + (1 + word_freq | participant)*. *time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < .05$, ** $p < 10^{-4}$

Predictor	**β**	**SE**	**z value**	**p**
word_freq2-1	-0.35	0.08	-4.27	10^{-4} **
time_point2-1	-0.01	0.08	-0.07	0.947
time_point3-2	0.21	0.09	2.48	0.013 *
time_point4-3	0.04	0.09	0.47	0.639
total_sleep_length_cent	0.09	0.14	0.65	0.515
SWS_min_cent	0.10	0.14	0.70	0.483
word_freq2-1:time_point2-1	-0.14	0.17	-0.86	0.390
word_freq2-1:time_point3-2	0.14	0.17	0.81	0.420
word_freq2-1:time_point4-3	-0.22	0.18	-1.22	0.223
word_freq2-1:total_sleep_length_cent	0.02	0.08	0.22	0.823
time_point2-1:total_sleep_length_cent	0.15	0.08	1.90	0.058
time_point3-2:total_sleep_length_cent	0.06	0.08	0.78	0.434
time_point4-3:total_sleep_length_cent	0.03	0.08	0.35	0.724
word_freq2-1:SWS_min_cent	0.08	0.08	1.02	0.307
time_point2-1:SWS_min_cent	0.01	0.09	0.13	0.896
time_point3-2:SWS_min_cent	-0.02	0.09	-0.23	0.818
time_point4-3:SWS_min_cent	-0.17	0.09	-1.83	0.067

word_freq2-1:time_point2-1:total_sleep_length_cent	-0.05	0.16	-0.29	0.775
word_freq2-1:time_point3-2:total_sleep_length_cent	0.14	0.16	0.83	0.407
word_freq2-1:time_point4-3:total_sleep_length_cent	-0.18	0.17	-1.08	0.279
word_freq2-1:time_point2-1:SWS_min_cent	-0.05	0.17	-0.26	0.792
word_freq2-1:time_point3-2:SWS_min_cent	0.07	0.18	0.41	0.685
word_freq2-1:time_point4-3:SWS_min_cent	0.33	0.18	1.80	0.072

Table S2C. Total sleep time and duration of SWS as predictors of performance in the generalisation task, in addition to time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: *accuracy ~ 1 + time_of_assessment * total_sleep_duration + time_of_assessment * SWS_duration + (1 | participant)*. *time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*.

* p < .05

Predictor	**β**	**SE**	**z value**	**p**
time_point2-1	0.15	0.10	1.42	0.514
time_point3-2	0.07	0.11	0.65	0.513
N3_min_cent	0.21	0.19	1.14	0.253
total_sleep_length_cent	0.09	0.19	0.48	0.630
time_point2-1:N3_min_cent	0.03	0.11	0.31	0.758
time_point3-2:N3_min_cent	0.05	0.11	0.47	0.642
time_point2-1:total_sleep_length_cent	0.27	0.11	2.52	0.012 *
time_point3-2:total_sleep_length_cent	-0.07	0.11	-0.63	0.530

Table S3A. Spindle density as a predictor of performance in the vocabulary task, in addition to word frequency and time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that Buildmer converged to: accuracy ~ 1 + word_freq * time_point * spindle_density_cent + (1 + word_freq | participant). *time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < .05$, ** $p < .01$, *** $p < 10^{-4}$

Predictor	**β**	**SE**	**z value**	**p**
word_freq2-1	-0.29	0.08	-3.87	10^{-4} ***
time_point2-1	0.20	0.08	2.56	0.010 **
time_point3-2	0.07	0.08	0.87	0.386
time_point4-3	-0.17	0.08	-2.11	0.036
spindle_density_cent	0.08	0.10	0.81	0.416
word_freq2-1:time_point2-1	0.18	0.16	1.13	0.261
word_freq2-1:time_point3-2	0.04	0.16	0.24	0.810
word_freq2-1:time_point4-3	-0.34	0.16	-2.14	0.033 *
word_freq2-1:spindle_density_cent	0.01	0.08	0.09	0.926
time_point2-1:spindle_density_cent	0.03	0.08	0.37	0.710
time_point3-2:spindle_density_cent	-0.08	0.08	-1.01	0.315
time_point4-3:spindle_density_cent	0.10	0.08	1.22	0.223
word_freq2-1:time_point2-1:spindle_density_cent	-0.02	0.16	-0.11	0.912
word_freq2-1:time_point3-2:spindle_density_cent	-0.21	0.17	-1.26	0.208
word_freq2-1:time_point4-3:spindle_density_cent	0.37	0.17	2.25	0.025 *

Table S3B. Spindle density as a predictor of performance in the trained inflections task, in addition to word frequency and time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that *Buildmer* converged to: *accuracy ~ 1 + frequency * time_point * spindle_density_cent* + *(1 + frequency | participant). time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*, and *4* for Day 9. In *word_freq*, *1* denotes frequent words, and *2* denotes infrequent words.

* $p < .05$, ** $p < 10^{-4}$

Predictor	β	SE	z value	p
word_freq2-1	-0.34	0.08	-4.23	$3x10^{-5}$ **
time_point2-1	0.00	0.08	-0.03	0.980
time_point3-2	0.21	0.09	2.42	0.016 *
time_point4-3	0.03	0.09	0.36	0.718
spindle_density_cent	-0.05	0.14	-0.39	0.700
word_freq2-1:time_point2-1	-0.14	0.17	-0.83	0.409
word_freq2-1:time_point3-2	0.13	0.17	0.75	0.454
word_freq2-1:time_point4-3	-0.20	0.18	-1.13	0.257
word_freq2-1:spindle_density_cent	-0.03	0.08	-0.37	0.710
time_point2-1:spindle_density_cent	-0.21	0.08	-2.50	0.013 *
time_point3-2:spindle_density_cent	0.05	0.08	0.59	0.555
time_point4-3:spindle_density_cent	0.10	0.09	1.11	0.266
word_freq2-1:time_point2-1:spindle_density_cent	-0.02	0.16	-0.15	0.883
word_freq2-1:time_point3-2:spindle_density_cent	-0.09	0.17	-0.52	0.600
word_freq2-1:time_point4-3:spindle_density_cent	0.14	0.17	0.82	0.415

Table S3C. Spindle density as a predictor of performance in the generalisation task, in addition to time of assessment.

Full accuracy results of the generalised logistic linear mixed-regression model that *Buildmer* converged to: *accuracy ~ 1 + time_point * spindle_density_NREM_cent + (1 | part_ID)*. *time_point* denotes time of assessment, with *1* for immediate, *2 for Day 2-morning*, *3 for Day 3-morning*.

Predictor	β	SE	z value	p
time_point2-1	0.14	0.10	1.31	0.191
time_point3-2	0.07	0.10	0.66	0.512
spindle_density_cent	-0.01	0.19	-0.06	0.954
time_point2-1:spindle_density_cent	0.16	0.11	1.51	0.131
time_point3-2:spindle_density_cent	-0.09	0.11	-0.87	0.383