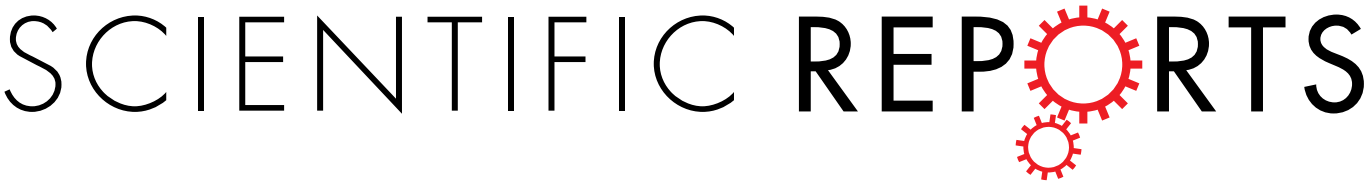

OPEN

Received: 31 May 2017
Accepted: 15 September 2017
Published online: 01 November 2017

Sleep Benefits Memory for Semantic Category Structure While Preserving Exemplar-Specific Information

Anna C. Schapiro[1], Elizabeth A. McDevitt[2], Lang Chen[3], Kenneth A. Norman[4], Sara C. Mednick[2] & Timothy T. Rogers[5]

Semantic memory encompasses knowledge about both the properties that typify concepts (e.g. robins, like all birds, have wings) as well as the properties that individuate conceptually related items (e.g. robins, in particular, have red breasts). We investigate the impact of sleep on new semantic learning using a property inference task in which both kinds of information are initially acquired equally well. Participants learned about three categories of novel objects possessing some properties that were shared among category exemplars and others that were unique to an exemplar, with exposure frequency varying across categories. In Experiment 1, memory for shared properties improved and memory for unique properties was preserved across a night of sleep, while memory for both feature types declined over a day awake. In Experiment 2, memory for shared properties improved across a nap, but only for the lower-frequency category, suggesting a prioritization of weakly learned information early in a sleep period. The increase was significantly correlated with amount of REM, but was also observed in participants who did not enter REM, suggesting involvement of both REM and NREM sleep. The results provide the first evidence that sleep improves memory for the shared structure of object categories, while simultaneously preserving object-unique information.

Semantic knowledge allows us to infer unobserved properties of newly encountered objects and events, requiring a mastery of both the coherent properties shared among conceptually related items and the individuating properties that distinguish them[1,2]. For instance, the concept *bird* arises when children learn that certain property sets — having wings, beaks, feathers, hollow bones, and the name "bird" — all co-occur together or not at all[2–4]. That is, the properties cohere and thus support judgments of similarity and generalization across the items that possess them. Conceptual systems must also, however, encode the properties that distinguish birds — the fact that penguins cannot fly, robins have red breasts, parrots are found in the tropics, and so on.

The Complementary Learning Systems (CLS) theory proposes a role for sleep in establishing long-term neocortical representations of such knowledge: The hippocampus rapidly stores new information while the organism is awake, which it then replays during sleep, allowing the slow-learning cortex to gradually incorporate the new information into existing knowledge structures[5]. Specialized subfields of the hippocampus rapidly bind arbitrary new information without interference using sparse, pattern-separated representations[6]. Neocortex, in contrast, employs more overlapping representations, facilitating extraction of the commonalities across individual elements and episodes. The hippocampus can judge commonalities in some situations, through recurrent dynamics at retrieval that allow activity to spread across episodic memories[7], or through associations learned in a particular subfield that has a moderately overlapping neural code[8]. However, cortical representations are more optimized for rapid and automatic identification of item similarities because highly overlapping representations there allow similar items to make more direct contact[2].

[1]Department of Psychiatry, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA. [2]Department of Psychology, University of California-Riverside, Riverside, CA, USA. [3]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA. [4]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA. [5]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA. Anna C. Schapiro and Elizabeth A. McDevitt contributed equally to this work. Correspondence and requests for materials should be addressed to A.C.S. (email: aschapir@bidmc.harvard.edu)

This view suggests that sleep should improve memory for coherent properties: As new learning examples become better integrated in the cortical system, representational overlap should promote cross-generalization amongst conceptually related items, boosting recall of these features. It may seem that idiosyncratic information would be lost in the overlapping cortical representations, and it is of course the case that we tend to forget the details of most of our memories over time. However, hippocampal replay during sleep may provide reminders about these details, helping cortex to represent this information as well, at least relative to a period of time awake.

A substantial literature now suggests that sleep benefits memory for both individual items and relational structure amongst items[9,10]. Prior research, however, has not assessed the relative fate and prioritization of these memory types during sleep, an understanding of which would provide important constraints on the mechanisms of consolidation. Assessing how sleep differentially impacts these memories requires matching strength of initial encoding, as forgetting rates vary with learning strength[11] and sleep tends to benefit more weakly encoded information[12–19], cf.[20,21].

We report two experiments using a novel semantic learning task, in which participants learned both coherent and individuating properties of 15 "satellite" objects organized into three categories. Satellites shared most features with other category members but also possessed unique individuating features. Learning as well as memory assessment for both feature types occurred via property inference[22], a task that captures the semantic system's primary function of inferring missing properties from partial information[1,2]. Frequency of exposure during learning was manipulated to ensure that (1) unique and shared properties were acquired equally well, but (2) the strength of this learning prior to sleep varied across the three categories, allowing us to assess how sleep prioritizes semantic memories with differential encoding strength. The paradigm thus employs a single canonical semantic task to assess memory for both coherent and individuating properties, with the degree of initial learning matched across property type but varied across categories.

In Experiment 1, participants in the Sleep condition learned at night and were tested before and after sleeping at home overnight, while participants in the Wake condition learned in the morning and were tested at the beginning and end of a day awake. In Experiment 2, participants learned and tested in the morning and tested again in the afternoon, either with or without a polysomnographically-recorded nap in between. This allowed us to compare behavioral results of an afternoon nap to overnight sleep and to assess the relationship of different sleep stages to changes in performance.

Methods

Participants. In Experiment 1, 111 members of the University of Wisconsin-Madison community (68 females, mean age =19.1 years, range =17–31 years) participated in exchange for course credit or monetary compensation. Data from 9 additional participants were excluded due to performing lower than 2 SD below average on the first session test (2 subjects) and procedural difficulties (7 subjects). The study protocol was approved by the Institutional Review Board at the University of Wisconsin-Madison, and methods were carried out in accordance with all guidelines and regulations.

In Experiment 2, 82 members of the University of California-Riverside community (49 females, mean age =19.9 years, range =18–34 years) participated in exchange for course credit or monetary compensation. Data from 11 additional subjects were excluded due to: napping for less than 20 min (3 subjects), falling asleep during the quiet wake period (7 subjects), and performing lower than 2 SD below average on the first session test (1 subject). Subjects reported having a regular sleep-wake schedule, which was defined as regularly going to bed no later than 2AM, waking up no later than 10AM, and getting at least 7 hours of total sleep per night on average. The Epworth Sleepiness Scale (ESS)[23] and the reduced Morningness-Eveningness Questionnaire (rMEQ)[24] were used to screen out potential subjects with excessive daytime sleepiness (ESS score >10) or extreme chronotypes (rMEQ <8 or >21). Experimental procedures were approved by the Human Research Review Board at the University of California at Riverside, and methods were carried out in accordance with all guidelines and regulations. Informed written consent was obtained from all participants in both experiments.

Stimuli. Participants learned about 15 novel "satellite" objects organized into three classes (Fig. 1a). Each satellite had a "class" name (Alpha, Beta, or Gamma) shared with other members of the same category, a unique "code" name (a well-formed nonword), and five visual parts. One of the satellites in each category is the prototype (shown on the left for each category in Fig. 1) — it contains all the prototypical parts for that category. Each of the other satellites has one part deviating from the prototype (which part deviates is different for each satellite in the category). Thus, each non-prototype shares 4 features with the prototype and 3 features with other non-prototypes from the same category. Exemplars from different categories do not share any features. Each satellite has *shared features*: the class name and the parts shared among members of the category; it also has *unique features*: the code name and the part unique to that satellite (except for the prototype, which has no unique part). Satellites were constructed randomly for each participant, constrained by this category structure.

Procedure. Participants learned about the satellites in two phases. In phase 1 (17 min mean duration), the satellites were introduced one by one (Fig. 1b). For each satellite, the class and code name were displayed followed by the satellite image. A box highlighted each visual feature in sequence to encourage participants to attend to each feature. Participants were then asked to recall the class and code names by clicking on one of three options given for each name. Next, participants used a point-and-click interface to try to reconstruct the satellite image from scratch. Icons representing the five part types were displayed on the right hand side of the screen, and when an icon was clicked, all the possible versions of that part were displayed in a row at the bottom of the screen. The participant could then click on one of the part versions on the bottom to add it to the satellite in the center of the screen. If the participant was too slow at this task (took longer than 15 s), or reconstructed the satellite incorrectly, a feedback screen would appear displaying the correct features.

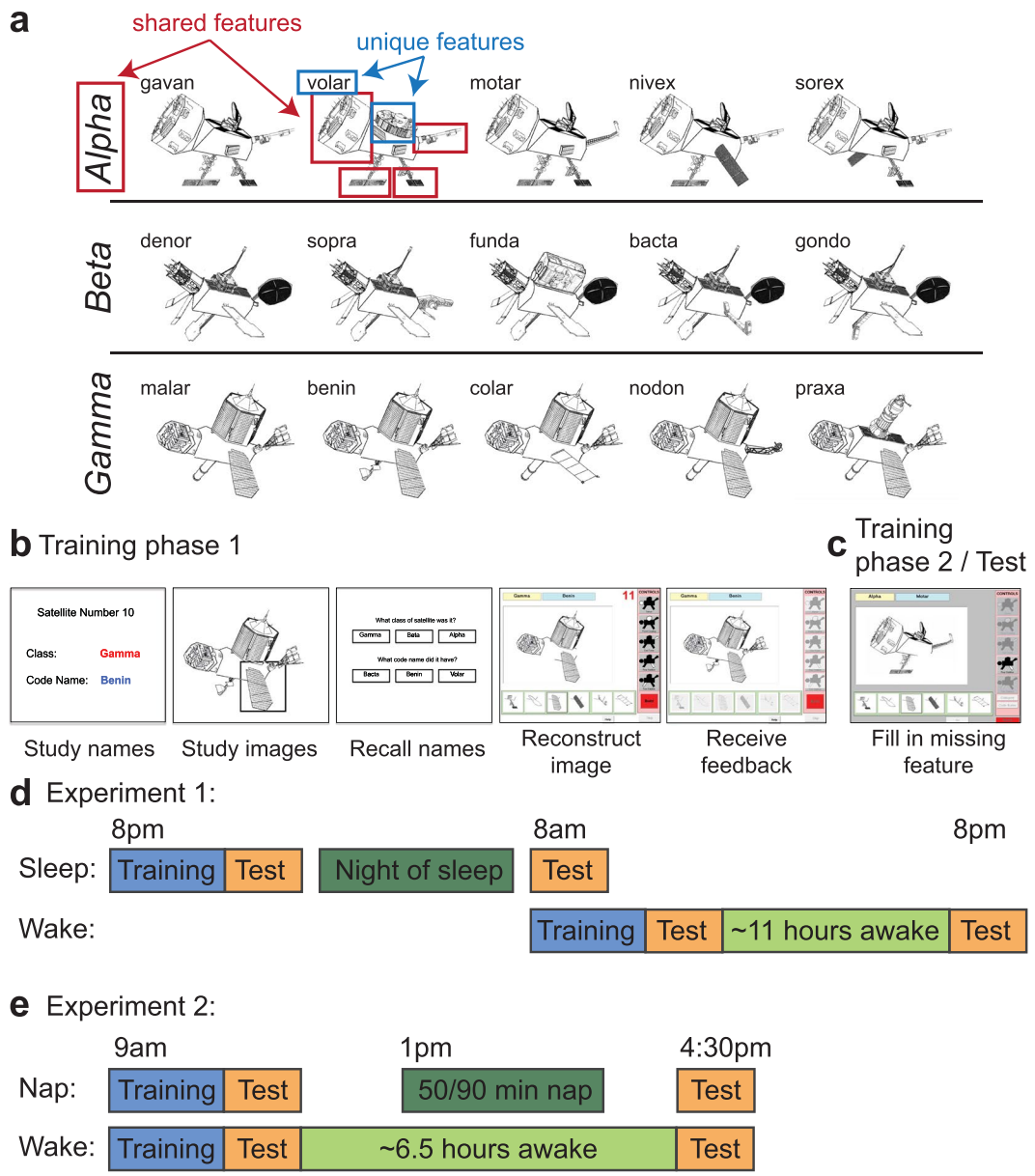

Figure 1. Stimuli and task. (**a**) Examples of satellite stimuli presented from the three classes Alpha, Beta, and Gamma, each labeled with unique code names. Satellites were built randomly for each participant, using the same category structure. Shared and unique features of one satellite are highlighted. (**b**) Training phase 1: one-by-one introduction to each satellite. (**c**) Training phase 2: participants attempt to fill in one missing feature of a satellite, receiving feedback. The test trials are the same but have two features missing. (**d**) Overview of protocol for Experiment 1. (**e**) Overview of protocol for Experiment 2. Time spans in (**d**) and (**e**) are not to scale.

In phase 2 of training (34 min mean duration) participants were shown a satellite with one feature missing (Fig. 1c), which could be one of the five visual features, the code name, or the class name (code and class name buttons were displayed along with the part icons on the right hand side of the screen, and when selected, displayed the corresponding name options in a row on the bottom). Using the same point-and-click interface, participants chose a feature amongst six options to complete the satellite. There was no response time deadline. If they chose the correct feature, they were told it was correct, and could move on to the next trial. If they chose an incorrect feature, they were shown the correct feature, and had to repeat the trial until they chose the correct feature. To assess effects of encoding strength, we varied the frequency of exposure of the different categories during this phase using a 1–2–3 ratio: 1/6 of all training items were from the low frequency (LF) category, 2/6 from the medium frequency (MF) category, and 3/6 from the high frequency (HF) category.

Remembering the shared properties of the satellites is easier than remembering the unique properties, as the shared properties are reinforced across study of all the satellites in the same class. The task was titrated in pilot testing to ensure that, at the end of training, participants performed equivalently at retrieving shared and unique properties of the satellites. To accomplish this, unique features were queried 24 times more frequently than shared

features. This exposure scheme results in the confounding of query frequency with feature type, which is necessary in order to avoid the confounding of performance level with feature type. This procedure thus leaves open the possibility that any differences between shared and unique feature item performance are due to differences in query frequency, as opposed to feature type per se. Phase 2 of training continued until the participant reached a criterion of 0.66 proportion of trials correct on a block of 32 trials, or until about an hour had passed (cutoff = 60 min in Experiment 1, cutoff = 75 min in Experiment 2).

Immediately after training, participants were tested by again filling in missing features of the satellites, now without feedback. The test phase had 51 trials, with two missing features per trial, which allowed us to collect more information per trial as well as provide less exposure to the correct features (to minimize learning during the test phase). Each satellite appeared twice in the test phase: once with its code name and its class name or one shared part tested, and once with two shared parts or one unique part and one shared part tested. The remaining 21 trials tested generalization to novel satellites. Novel satellites were members of the trained categories but had one novel feature or a novel combination of features (a prototype from one category with one or two prototypical features swapped in from a different category). The queried feature for novel items was always a shared feature (class name or shared part). Test trials were presented in a random order.

Participants completed the same set of test items (in a different random order) after a delay. They were not told in the first session that there would be a second memory test. The Karolinska Sleepiness Scale (KSS)[25], which assesses state sleepiness/alertness on a scale of 1 (extremely alert) to 9 (very sleepy), was completed at the end of each session.

Experiment 1-specific procedure. In Experiment 1, 12 hours elapsed between the two sessions (Fig. 1d). Participants in the Sleep condition (n = 61) began the first session around 8 pm, and participants in the Wake condition (n = 50) began the first session around 8am. Participants were given no instructions about their activities between sessions. At the end of the second session, subjects in both groups filled out a questionnaire asking how long they slept between sessions. The KSS was not collected for 22 subjects due to procedural error.

Experiment 2-specific procedure. Subjects participating in this polysomnography (PSG) study underwent stricter sleep screening and procedures. Subjects were instructed to keep a regular sleep schedule and attempt at least 7 hours of sleep per night. Adherence to the sleep schedule was tracked with daily sleep diaries. Additionally, subjects wore an actigraph wrist monitor (Actiwatch Spectrum, Respironics) the night immediately prior to the study day. Subjects were asked to refrain from consuming caffeine, alcohol, and all stimulants for 24 hours prior to and including the day of the study. Heavy caffeine users (>240 mg per day) were not enrolled to exclude the possibility of significant withdrawal symptoms during the experiment.

Subjects arrived between 8:30am–9am (Fig. 1e). Before proceeding with the experiment, an experimenter checked each subject's actigraphy data to verify adherence to the sleep schedule the night prior. Session 1 began at approximately 9am.

Upon completion of Session 1, subjects were randomly assigned to one of four groups: Active Wake (AW), Quiet Wake (QW), 50 min nap, or 90 min nap. The AW group (n = 22) carried out their normal daily activities outside of the lab, but were instructed to abstain from exercise and napping. Wakefulness in the AW group was monitored with the actigraph wrist monitors. Subjects in the QW and Nap groups had electrodes attached for standard PSG. At 1 pm, the QW group (n = 20) commenced listening to short stories on an iPod for 60 min while sitting in a quiet, dark room with PSG monitoring to make sure they did not fall asleep. During QW sessions, an experimenter woke subjects at the first sign of Stage 1 sleep. Subjects in the two Nap groups were given a nap opportunity between 1–3PM. If a subject spent more than 30 consecutive minutes awake during the nap window then the nap was ended. Otherwise, an experimenter woke the subject after he or she had obtained the desired amount of total sleep time (50 min or 90 min). Given that shorter naps tend to have less REM sleep than longer naps, the use of these two durations increased the likelihood of having naps with and without REM sleep. Post-hoc sleep stage scoring was used to place subjects into a REM group (n = 23, naps contained more than one minute of REM sleep) or non-REM (NREM, n = 17) after completion of the experiment. Ten subjects who were assigned to the 50 min nap had enough REM sleep to qualify for the REM condition, and one subject who was assigned to the 90 minute nap did not have enough REM sleep to qualify for the REM condition. The Session 2 test occurred for all groups between 4:30–5 pm.

Polysomnography. PSG data were collected using Astro-Med Grass Heritage Model 15 amplifiers and Grass Gamma software. Eight scalp electroencephalogram (EEG) and two electrooculogram (EOG) electrodes were referenced to unlinked contralateral mastoids (F3/A2, F4/A1, C3/A2, C4/A1, P3/A2, P4/A1, O1/A2, O2/A1, LOC/A2 and ROC/A1), and two electromyogram electrodes were attached under the chin to measure muscle tone. PSG data were digitized at 256 Hz and visually scored in 30 s epochs according to the sleep staging criteria of Rechtschaffen and Kales[26]. Sleep architecture variables included percentage of the nap spent in Stage 1, Stage 2, slow wave sleep (SWS), and rapid eye movement (REM) sleep.

EEG data were preprocessed and analyzed using BrainVision Analyzer 2.0 (BrainProducts, Munich Germany) and Matlab. EEG data were bandpass filtered between 0.3 and 35 Hz, and all epochs with artifacts and arousals were identified by visual inspection and rejected. Sleep spindles were automatically detected during Stage 2 and SWS using a wavelet-based algorithm[27]. Following spindle detection, spindle densities were calculated by dividing the number of discrete spindle events by minutes spent in each sleep stage at each scalp EEG electrode site. Data for an individual channel were excluded if the channel was determined to be unreliable.

Multiple comparisons correction. We used False Discovery Rate (FDR)[28] correction for multiple comparisons where noted.

Data availability. The data analyzed in this study are included as Supplementary Information.

Results

Experiment 1. *Sleepiness survey.* The KSS scores did not differ across Sleep and Wake conditions for the first session (mean Sleep = 5.67; mean Wake = 6.15; $t[86] = 1.406$, $p = 0.163$) nor second session (mean Sleep = 5.46; mean Wake = 4.88, $t[86] = 1.499$, $p = 0.138$), suggesting that there were no alertness differences between groups due to time of day.

Training. Participants trained for an average of 171.55 trials (SD = 84.39), including repetition trials for incorrect choices. The average proportion correct on the last training block was 0.686 (SD = 0.150).

Test performance. For proportion correct on the first test, we found no differences between Wake and Sleep groups in unique, shared, or novel feature types (ps > 0.175; Supplementary Fig. S1), indicating no bias due to time of day. We also found no differences between first session unique and shared performance within each of the two groups (ps > 0.164; Supplementary Fig. S1), demonstrating a successful matching of performance on these feature types. Novel items, which we did not attempt to match in performance, were worse than unique and shared features in the first test: $t[60] = 1.586$, $p = 0.118$ for the Sleep group; $t[49] = 2.987$, $p = 0.004$ for the Wake group.

The frequency manipulation had a robust effect on Session 1 performance (Supplementary Fig. S1), which was most pronounced for unique items (LF mean = 0.406, SD = 0.245; MF mean = 0.530, SD = 0.235; HF mean = 0.659, SD = 0.247), followed by shared items (LF mean = 0.487, SD = 0.229; MF mean = 0.535, SD = 0.217; HF mean = 0.613, SD = 0.247), and then novel items (LF mean = 0.455, SD = 0.208; MF mean = 0.472, SD = 0.198; HF mean = 0.531, SD = 0.219). Individual subject slopes across the three frequency levels were significantly greater for unique than shared ($t[110] = 3.983$, $p = 0.0001$), and marginally greater for shared than novel ($t[110] = 1.687$, $p = 0.095$).

To assess change in proportion correct from the first to second session, we ran a three way ANOVA on the Session 2 − Session 1 performances, with Sleep vs. Wake groups as across-subject factor and frequency and feature type as within-subject factors. There was a main effect of group, with greater (i.e., more positive) change for the Sleep group than for the Wake group ($F[1,109] = 9.802$, $p = 0.002$), a main effect of feature type ($F[2,218] = 5.913$, $p = 0.003$), with unique features improving less than shared and novel, and a main effect of frequency ($F[2,218] = 3.090$, $p = 0.047$), with lower frequencies faring better (i.e., showing less forgetting) than higher frequencies. Finally, there was a group by feature type interaction ($F[2,218] = 4.074$, $p = 0.018$), driven by there being no difference between Sleep and Wake in the novel item features ($t[109] = 0.235$, $p = 0.815$) but a difference between groups in both unique ($t[109] = 3.098$, $p = 0.003$) and shared features ($t[109] = 3.267$, $p = 0.002$). The magnitude of the Sleep vs. Wake difference for unique vs. shared features was not different, collapsing across frequency ($t[109] = 0.303$, $p = 0.763$). These results suggest that a night of sleep is much more beneficial than a day awake for memory for unique and shared features of the satellites.

Notably, there was a reliable above-baseline improvement in shared feature memory in the Sleep group, collapsing across frequency (Fig. 2; $t[60] = 2.093$, $p = 0.041$; after FDR correction across three feature types, $p = 0.122$). This provides evidence that subjects have an improved understanding of the shared category structure after sleeping. Average change in performance for the Sleep group was marginally better for shared than unique features ($t[60] = 1.688$, $p = 0.097$). Breaking the data down by category frequency, the Sleep group was not significantly different from zero in any of the nine frequency-specific conditions (ps > 0.131).

The Wake group had reliable forgetting of both unique and shared features, collapsing across frequency (unique: $t[49] = 4.109$, $p = 0.0002$, FDR-corrected $p = 0.0004$; shared: $t[49] = 2.482$, $p = 0.017$, FDR-corrected $p = 0.026$). Breaking the data down by category frequency, there was significant forgetting in MF unique features ($t[49] = 3.805$, $p = 0.0004$, FDR-corrected $p = 0.004$), HF unique features ($t[49] = 3.452$, $p = 0.001$, FDR-corrected $p = 0.005$), and HF shared features ($t[49] = 3.105$, $p = 0.003$, FDR-corrected $p = 0.010$). No other conditions differed significantly from zero (ps > 0.173). Forgetting was reliably greater in MF and HF unique features than in LF unique features (MF: $t[49] = 2.117$, $p = 0.039$; HF: $t[49] = 2.017$, $p = 0.049$). Forgetting was also reliably greater in HF shared features than LF shared features ($t[49] = 2.038$, $p = 0.047$).

Verbal vs. visual information. Each feature type had verbal and visual subtypes: Unique features could be a visual part or a verbal code name, shared features could be a visual part or a verbal class name, and novel item features could be a visual part or a verbal class name. To assess whether verbal vs. visual type had any effect, we re-ran the ANOVA with verbal vs. visual as an additional within-subject factor and found a large main effect of this factor ($F[1,109] = 18.794$, $p = 0.00003$), with verbal information remembered better than visual across the delay. There were no interactions between this factor and group (or any other factors), however, suggesting that sleep does not have a different effect depending on whether the feature is learned verbally vs. visually.

Learning within the test sessions. We designed the tests to minimize the potential for learning during the test phases. To verify that no learning occurred, we ran an ANOVA on the within-session performances, with frequency, feature type, and first vs. second half of the test phase as factors. There were no main or interaction effects for first vs. second half of the test phase (ps > 0.260), indicating no evidence of learning. Mean proportion correct in the first half of the test phases was 0.520 vs. 0.516 in the second half.

Nap survey. Out of 50 subjects in the Wake group, 15 reported taking a nap between the first and second session. The mean nap length was 64.0 minutes (median = 60, SD = 35.36, range = 15–135). These subjects did not differ from those who did not nap in their change from first to second session performance for any feature type at any

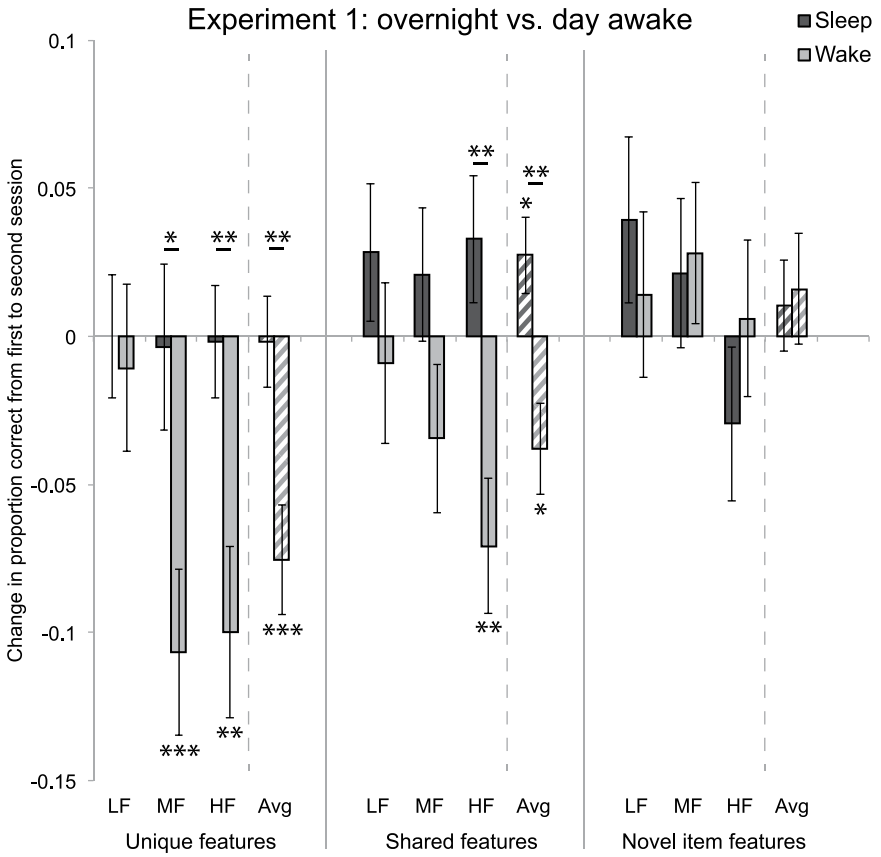

Figure 2. Experiment 1 results. Change in proportion correct from first to second session for unique features, shared features, and novel item features. For each feature type, results are shown for low frequency (LF), medium frequency (MF), and high frequency (HF) category, as well as the average (Avg) across categories. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, uncorrected. Asterisks above horizontal lines show significant differences between conditions; asterisks without bars indicate where conditions differ from zero. Error bars denote ±1 SEM.

frequency level ($ps > 0.138$), suggesting that any benefit of the nap was swamped by the effect of a much longer period of time awake. Including subjects who napped in the Wake group might have been expected to reduce the magnitude of the effects, but Cohen's d values were slightly stronger (more negative) for unique (nap $d = -0.564$, no nap $d = -0.356$) and shared features (nap $d = -0.310$, no nap $d = -0.182$) in the group that napped, collapsing across frequency (for novel item features, nap $d = 0.140$, no nap $d = 0.075$). We therefore collapsed across these two subgroups for all reported analyses.

Discussion. The overall picture from Experiment 1 is that a night of sleep promotes retention of the unique features of category members and improves memory for shared features, whereas both of these feature types are forgotten across a day awake. The forgetting effect was larger for higher-frequency items, which may in part reflect a proportional forgetting function (i.e., the better memory is initially, the larger forgetting will be). The Sleep group resisted this forgetting effect, across all frequency levels. There were no changes or group effects in generalization to novel item features (see General Discussion for possible explanations).

We next ran a nap variant of the paradigm with polysomnography, allowing us to explore the potential influence of sleep features and the effects of a shorter sleep/wake period, with tests at matched times of day across groups.

Experiment 2. *Sleepiness survey.* In the first session, there was no difference in KSS scores between the groups that did sleep versus those that did not (mean Sleep groups = 4.025; mean Wake groups = 4.357; $t[80] = 0.803$, $p = 0.425$). In the second session, the groups that slept reported being less sleepy than the groups that did not (mean Sleep groups = 2.700; mean Wake groups = 4.381; $t[80] = 4.524$, $p < 0.001$). However, KSS scores in the second session did not correlate, in either group, with second session performance or change from first to second session performance for any feature type ($ps > 0.123$).

PSG data. A summary of the PSG data is shown in Table 1. By design, the REM group had greater total sleep time ($t[38] = 4.552$, $p = 0.0001$) and more minutes spent in Stage 2 ($t[38] = 2.148$, $p = 0.038$) and REM sleep ($t[38] = 7.107$, $p < 0.0001$). There were no differences in minutes spent in Stage 1 ($t[38] = 0.790$, $p = 0.435$) or slow wave sleep ($t[38] = 0.417$, $p = 0.679$). REM nappers had greater sleep efficiency (total sleep time divided by time in bed; $t[38] = 2.709$, $p = 0.010$), indicating that they spent a greater proportion of the nap opportunity asleep.

	NREM naps	REM naps
TST (min)***	52.9 (11.8)	76.8 (19.1)
SE (%)*	78.9 (16.5)	89.7 (8.5)
Minutes		
Stage 1	6.5 (6.2)	5.3 (3.1)
Stage 2*	27.0 (16.3)	36.9 (12.9)
SWS	19.3 (13.4)	20.9 (11.2)
REM***	0.1 (0.2)	13.7 (7.9)
Percent (% TST)		
Stage 1	13.2 (14.7)	7.3 (4.2)
Stage 2	49.4 (21.1)	48.1 (11.8)
SWS	37.3 (25.9)	27.2 (12.1)
REM***	0.1 (0.3)	17.4 (8.3)

Table 1. Sleep architecture descriptives. *Note*. Values are M (SD). TST = total sleep time; SE = sleep efficiency; SWS = slow wave sleep; REM = rapid eye movement. Stars indicate significant differences between NREM and REM groups. *$p < 0.05$, ***$p < 0.001$.

Training. Participants trained for an average of 199.96 trials (SD = 108.23). Average performance on the last training block was 0.748 (SD = 0.060).

Test performance. For proportion correct on the first test, we found no differences between QW, AW, NREM, or REM groups in unique, shared, or novel feature types ($ps > 0.150$), indicating no bias across groups before the delay. We also found no differences between first session unique and shared performance within any of the four groups ($ps > 0.172$). Novel items, which we did not attempt to match in performance, were worse than unique and shared features in the first test: $t[16] = 4.447$, $p < 0.001$ for NREM; $t[22] = 4.803$, $p < 0.001$ for REM; $t[21] = 3.095$, $p = 0.006$ for AW; $t[19] = 1.829$, $p = 0.083$ for QW.

Looking at the change in proportion correct from the first to second session, we found no differences between QW and AW, nor between NREM and REM groups, for any frequency level or feature type ($ps > 0.076$). We therefore collapsed the groups into Nap (NREM and REM) and Wake (QW and AW) for further analyses, except when assessing the contribution of sleep stages.

The frequency manipulation had a robust effect on Session 1 performance (Supplementary Fig. S2), which was again most pronounced for unique items (LF mean = 0.391, SD = 0.200; MF mean = 0.577, SD = 0.199; HF mean = 0.745, SD = 0.189), followed by shared items (LF mean = 0.537, SD = 0.198; MF mean = 0.603, SD = 0.202; HF mean = 0.646, SD = 0.207), and then novel items (LF mean = 0.477, SD = 0.180; MF mean = 0.460, SD = 0.192; HF mean = 0.494, SD = 0.209). Individual subject slopes across the three frequency levels were significantly greater for unique than shared ($t[81] = 6.690$, $p < 0.0001$), and significantly greater for shared than novel ($t[81] = 2.629$, $p = 0.010$).

The Nap and Wake nap groups did not differ from the Sleep and Wake groups in Experiment 1 in first session performance, for any frequency or item type ($ps > 0.0667$), but the pattern of change over time was different. We again ran a three way ANOVA on Session 2 − Session 1 performance, with Nap vs. Wake groups as across-subject factor and frequency and feature type as within-subject factors. There was a main effect of feature type ($F[2,160] = 10.450$, $p = 0.00005$), a feature type by frequency interaction ($F[4,320] = 3.104$, $p = 0.016$), and, critically, a feature type by frequency by group interaction ($F[4,320] = 3.104$, $p = 0.016$). This indicates that there was no overall effect of Nap versus Wake, nor differences between the groups for unique, shared, and novel feature types, but rather that the groups differed in certain frequencies and feature types. In particular, the Nap group improved more than the Wake group on LF shared features ($t[80] = 2.846$, $p = 0.006$; with FDR correction across 3 feature types, 3 frequencies, $p = 0.051$; Fig. 3). This may reflect a prioritization of the category that is least well learned.

Notably, as in Experiment 1, this change in LF shared feature performance reflected an above-baseline improvement for subjects who slept — the Nap group performed better on LF shared features after the nap than before ($t[39] = 2.888$, $p = 0.006$). There was also a difference between the Nap and Wake groups for HF shared features ($t[80] = 2.583$, $p = 0.012$; with FDR correction, $p = 0.052$), with Wake better than Nap, though this did not reflect above baseline improvement for the Wake group ($t[41] = 1.105$, $p = 0.275$). The feature type by frequency interaction reflects the fact that there was more forgetting, across both groups, for higher frequency unique and shared features (but not for novel item features).

Verbal vs. visual information. We again tested for effects of verbal vs. visual feature type by re-running the ANOVA with verbal vs. visual as an additional within subject factor. We found a main effect ($F[1,80] = 4.469$, $p = 0.038$), with verbal information again remembered better than visual across the delay. There were again no interactions between this factor and group (or any other factors).

Learning within the test sessions. To verify that no learning occurred during the test sessions in Experiment 2, we again ran an ANOVA on within-session performances, with frequency, feature type, and first vs. second half

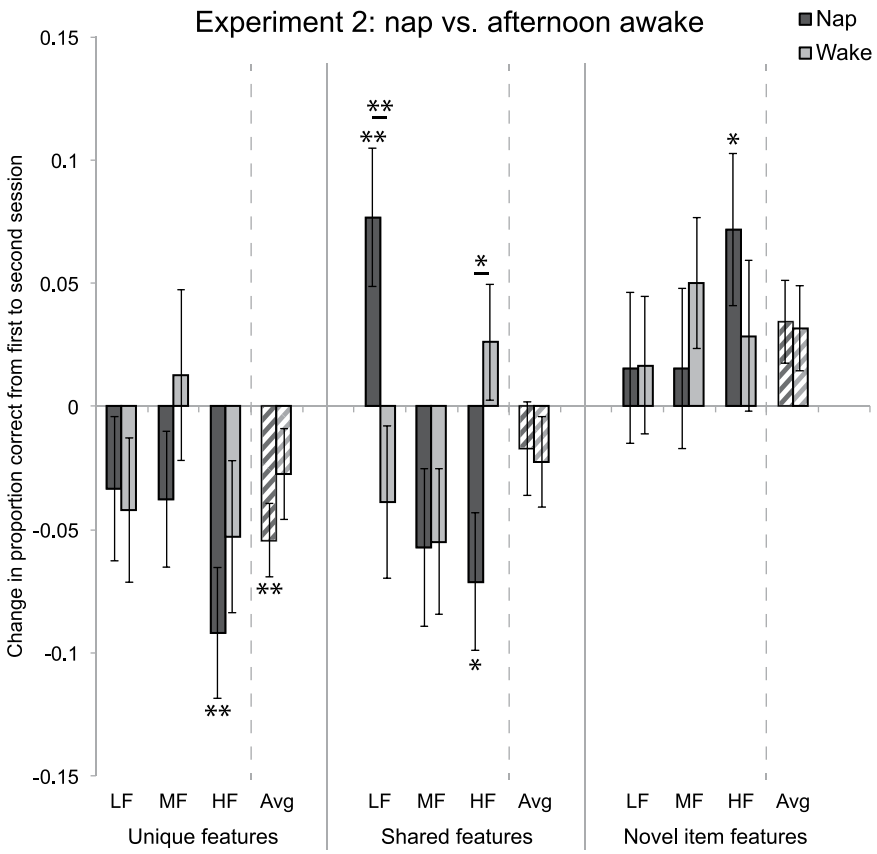

Figure 3. Experiment 2 results. Same structure as in Fig. 2. *$p < 0.05$, **$p < 0.01$, uncorrected.

of the test phase as factors. We again found no main or interaction effects for first vs. second half of the test phase ($ps > 0.428$). Mean proportion correct in the first half of the test phases was 0.545 vs. 0.541 in the second half.

Sleep features. To assess the potential contribution of different sleep stages to the improvement in shared features in the LF category, we first looked at performance separately for the NREM and REM groups (Fig. 4a). The NREM group improved significantly from Session 1 to 2 ($t[16] = 2.954$, $p = 0.009$), suggesting that NREM sleep alone is sufficient for this change in performance. The REM group was no different than the NREM group ($t[38] = 0.525$, $p = 0.603$), but was not itself reliably above baseline ($t[22] = 1.602$, $p = 0.123$).

We next used linear regression to assess whether the amount of time in each sleep stage was related to change in performance for LF shared features. This approach allows assessment of the contribution of each sleep stage controlling for time spent in other stages and for total sleep time. We also included Session 1 performance as a predictor, to control for any influences of initial performance. Visual inspection of the data suggested a nonlinear effect of REM sleep, so we included a quadratic term for time in REM. In this model, sleep stages explained 50.5% of the variance in behavioral change ($F[6,33] = 5.601$, $p = 0.0004$). The linear predictor for minutes spent in REM was marginally significant ($p = 0.067$) whereas the quadratic predictor was highly significant (Fig. 4b; $p = 0.002$). There were no other significant predictors ($ps > 0.171$). This pattern suggests that a small amount of REM is actually worse than having no REM at all, and that NREM sleep followed by sufficient REM is optimal. Within the REM group, there was a significant correlation between performance improvement and minutes of REM sleep ($r = 0.592$, $p = 0.002$) as well as proportion of time in REM sleep ($r = 0.656$, $p = 0.001$). We also performed a median split analysis on number of minutes spent in REM sleep, which divided the REM group into subjects getting 1–12 minutes and those getting 13–37 minutes (Fig. 1a). Subjects who get less REM showed a numerical decrement in performance, whereas those with more REM performed numerically better than the NREM group (difference between less REM and more REM: $t[21] = 3.212$, $p = 0.004$). This pattern again suggests that a small amount of REM is worse than having no REM at all (less REM vs. NREM: $t[26] = 2.966$, $p = 0.006$).

To test whether the relationship between REM sleep and behavioral improvement was specific to the LF shared features, which would provide converging evidence for a prioritization in processing of the LF category during sleep, we ran the same regression for MF and HF shared features. In both cases, the overall models were not significant ($ps > 0.457$), nor were any of the individual predictors ($ps > 0.194$). The correlation between proportion of time spent in REM and performance improvement was reliably greater for LF than MF ($p = 0.03$) or HF ($p = 0.008$) shared features, suggesting that the nap indeed focused on processing of shared features in the LF category (Fig. 4c,d).

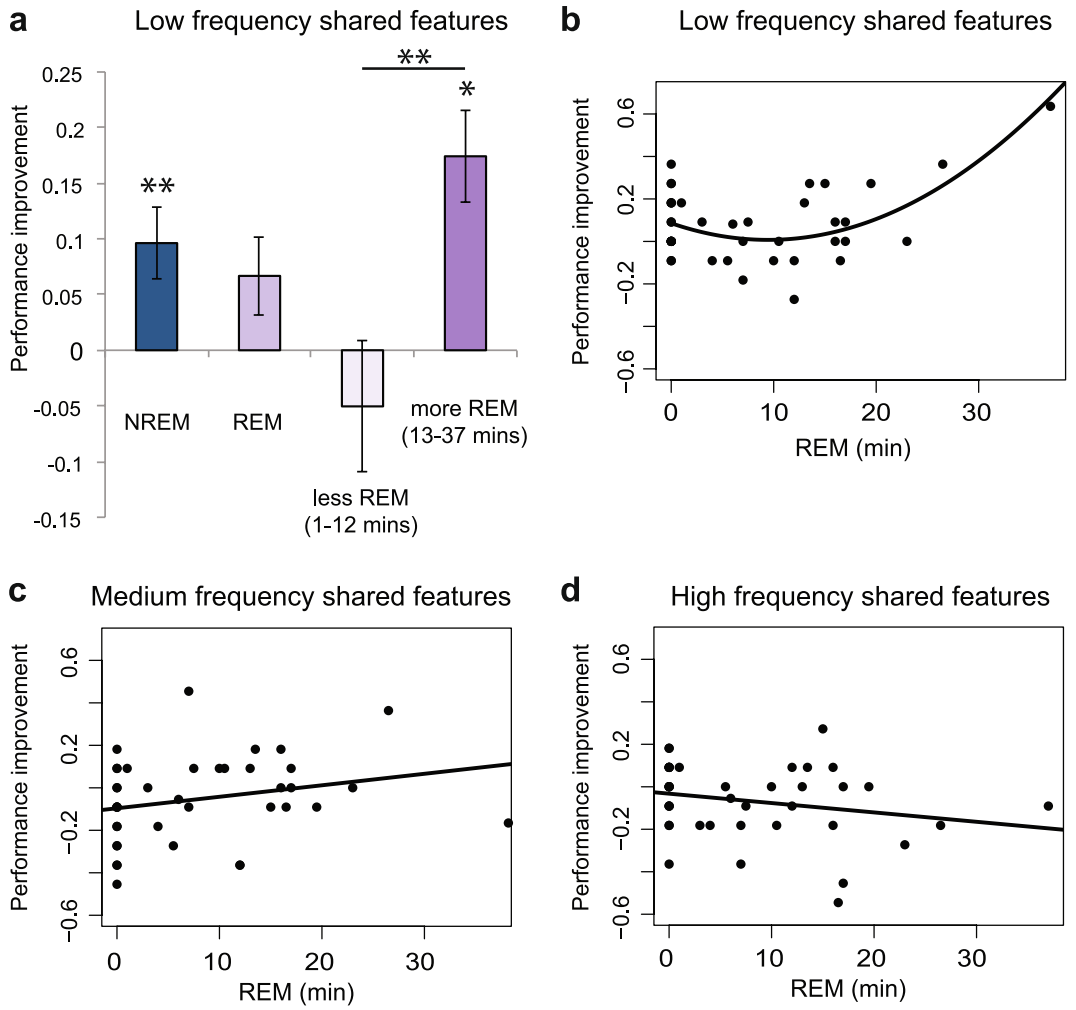

Figure 4. Relationship between NREM and REM sleep and behavioral change. (**a**) Improvement in low frequency shared features for the NREM group, the REM group, and for REM groups on either side of a median split by number of minutes in REM. (**b**–**d**) Relationship between minutes of nap spent in REM and performance improvement for LF, MF, and HF shared features. *$p < 0.05$, **$p < 0.01$.

We next performed six exploratory regressions for LF, MF, and HF unique and novel features. None of the overall models ($ps > 0.203$) nor individual predictors ($ps > 0.069$) were significant. We performed further exploratory analyses assessing whether spindle density in Stage 2 or SWS at any channel correlated with improvement for any of the three feature types, at any of the three frequency levels. While there were reliable uncorrected correlations, none survived FDR correction ($ps > 0.138$).

Discussion. Over a short period of sleep, we again found an above-baseline improvement in shared feature memory, but only for the LF category — the category that starts off with the poorest performance. This suggests a potential prioritization of the weakest category early in a sleep period.

There were no differences between NREM and REM groups in terms of behavioral change, nor between AW and QW, in contrast to prior studies[16]. When considering the PSG data, the group with only NREM sleep showed reliable improvement in LF shared features, suggesting that NREM sleep is sufficient to promote better category understanding, and time in REM sleep was associated with the improvement, suggesting that REM is also doing relevant processing.

For unique and shared features, as in Experiment 1, memory declined more for the higher frequency categories, which start off better learned. In this case, the frequency-influenced forgetting occurred for both Nap and Wake groups (with the puzzling exception of HF shared features in the Wake group). We again found no notable changes in novel item features.

General Discussion

Effective consolidation of new semantic learning must simultaneously encompass both object-specific information and the conceptual structure existing across objects. Experiment 1 showed that a night of sleep indeed benefits both knowledge types: Whereas shared and unique features showed forgetting (especially for well-learned

items) over wake, memory was preserved for unique features and improved for shared features over sleep regardless of initial encoding strength. Experiment 2 found memory improvements for lower-frequency shared features in a nap paradigm, with polysomnography suggesting that both NREM and REM phases contributed to these gains. In this section we consider the implications of these results for our understanding of semantic knowledge consolidation.

Enhancement versus maintenance of new learning. The existing sleep literature has suggested that, though sleep may enhance non-declarative learning, it only serves to prevent or decrease forgetting of declarative memory, perhaps just passively protecting against interference[29,30], cf.[31]. Contrary to this view, we observed above-baseline improvement from Session 1 to Session 2 in memory for shared features in a declarative memory task, with no evidence of learning within the sessions, suggesting an active consolidation process[32]. What accounts for the difference? Prior studies of declarative memory in the sleep literature have mainly used episodic memory tasks, requiring arbitrary associations (e.g., word pairs or object locations). Our results suggest that the division between what is maintained versus enhanced by sleep may not be about declarative vs. nondeclarative information, but instead about arbitrary versus structured information. Our unique feature condition, in which we found a prevention of forgetting overnight, is more episodic-like, in the sense that the associations are arbitrary, and may correspond to the kinds of processing observed in prior episodic memory sleep studies. Our shared feature condition, like prior studies of non-declarative memory, would be expected to benefit from a stronger cortical representation. Furthermore, other studies of structured information have also found above-baseline improvement[18,33–36]. The current data suggest that, for new semantic learning, sleep promotes maintenance of arbitrary information and simultaneous enhancement of memory for shared structure.

Prioritization of weak memories. In Experiment 2, which had shorter sleep and wake intervals between tests, we again saw a true improvement in shared feature memory in the Sleep group, but only for the category that had been exposed at the lowest frequency during training, with forgetting occurring for the high frequency category. The difference between Nap and Wake groups in the low frequency shared features, which was the most robust group difference observed, was marginal after correcting for multiple comparisons ($p = 0.051$). Our conclusions regarding Experiment 2 are thus tentative, and more research will be needed to clarify the pattern of change over a nap. However, the finding is consistent with an accumulating literature suggesting that weaker memories are prioritized during sleep[12–19], cf.[20,21]. Combined with Experiment 1, the results suggest the possibility that sleep first prioritizes the information most in need of help, allowing some forgetting of non-prioritized information, and then moves on to that other information. Another way of conceptualizing this result is that a nap results in a normalization of the pre-nap frequency effect, where the low frequency category was worst remembered and the high frequency category best remembered. The mechanism by which weaker memories may be tagged for early prioritization is an important topic for future research. Subjects in these studies did not receive feedback during the test phase, so tagging would have had to rely on some internal sense of performance accuracy.

Experiment 2 also suggests that initial strength of encoding is not the only factor determining the subsequent fate of memories over sleep or wake. Low frequency unique features did not benefit from a nap despite being more weakly encoded than low frequency shared features. That is, sleep only benefited weakly-learned properties shared amongst category members. Thus, the degree to which sleep appears to prioritize some memories over others depends not only on strength of initial learning and amount of sleep, but, again, on the degree to which the information is structured.

Forgetting of unique features differs for nap versus full night sleep. Experiments 1 and 2 both found that memory for unique features declined over wake and that low frequency unique features were preserved over sleep. The pattern was very different, however, for high frequency unique features: A full night's sleep produced good retention of unique features regardless of frequency, while a short nap produced significant forgetting of high frequency unique features (Experiment 1 vs. 2 sleep groups for high frequency unique features: $t[99] = 2.674$, $p = 0.009$). The pattern makes sense if early sleep prioritizes weaker memories — without active processing during the nap, the stronger items are forgotten as if the participant had stayed awake, but more sleep (or more NREM-REM cycles) provides opportunities for rescue. The combined results are thus consistent with the possibility that unique, in addition to shared features, are processed in an active way during sleep.

These and other differences between the nap and overnight studies suggest that for this paradigm, unlike others[37], a nap is not equivalent to a full night of sleep — a shorter sleep opportunity in a rich object category learning paradigm may only allow focus on items most in need of help. The fact that subjects who took a nap during the day in the Experiment 1 Wake condition did not show a different pattern of behavior from the other Wake subjects additionally suggests that the effects of a nap may not be long-lasting in this paradigm.

Active gains occur in both NREM and REM sleep. Experiment 2 found that NREM sleep alone improves memory for low frequency shared features. The REM group showed numerically smaller benefits, but the amount of REM sleep was strongly positively correlated with improvement — a relationship specific to low frequency shared features. Considering all nappers in one regression, there was a quadratic relationship between time in REM and performance improvement. This suggests that NREM plus sufficient REM sleep most strongly benefits knowledge of category structure, but that NREM plus a small amount of REM can degrade such knowledge. Both NREM and REM sleep have previously been associated with memory for structured information. The amount of NREM sleep correlates with associative inference[38] and statistical learning of tone sequences[33], and the amount of REM sleep is associated with improved performance on the remote associates task[39]. Both sleep stages have been associated with learning a hidden linguistic rule[40] and probabilistic learning on the weather prediction

task[18,41]. NREM sleep may be the time when the hippocampus replays new information to cortex, while REM helps to further stabilize the new information in cortical networks[42]. Both of these processes would be expected to benefit understanding of shared structure. The possibility that conceptual structure benefits most from hippocampal replay followed by cortical stabilization is a hypothesis that we hope to address in future work.

Sleep does not influence generalization in the property inference task. In neither experiment did we observe differences between sleep and wake groups in generalization to novel satellites. This was surprising, as a better understanding of shared category structure would be expected to lead to better ability to generalize[2], and prior studies have found benefits of sleep on generalization, including enhancement from pre-sleep levels[34–36,43–46], cf.[47,48]. It is possible that a longer period of time would need to elapse before seeing generalization in this paradigm[49], or that our assessment was not sensitive enough: Generalization was better than chance but reliably worse than memory for unique and shared properties. It is also possible that there was variance across participants in the strategies used for making judgments about never-before-seen features that hindered our ability to detect generalization.

If the lack of a sleep benefit for generalization is not due to insensitivity or variance in strategy use, how can we interpret this null result? One possibility is that the improvement in memory for shared features in the sleep group does not reflect cortical learning of category structure, but instead improved hippocampal memory for individual satellites. Shared feature memory can in principle be supported by hippocampal episodic memory (in addition to cortical structure learning), and thus improved memory for individual satellites could be supported by strengthened hippocampal item traces without new cortical learning. Under this account, shared features exhibit more robust improvement than unique features because cortical category structure (acquired during initial wake learning, but not boosted by sleep) would serve to selectively amplify hippocampal recall of shared (vs. unique) features. Adjudicating between this account (hippocampal strengthening without cortical learning during sleep) and accounts that posit cortical structure learning during sleep will require further work carefully assessing generalization in this paradigm. Of note, a paradigm looking at systematic versus arbitrary mappings in an artificial language paradigm also found a sleep-dependent benefit for both types of features but not for generalization[50]. However, performance was not matched or measured for the feature types prior to sleep, so it is unknown whether sleep was boosting these memories above baseline.

Other related work. There have been a few prior studies finding effects of sleep on semantic memory[45,51], but they have focused on integrating new information with existing semantic memory networks, not learning an entirely novel domain, as in our study. One study that did look at novel conceptual learning found retention of memory for category exemplars as well as retention of ability to generalize to novel exemplars and never-seen prototypes across a night of sleep, but not a day awake, in a dot pattern categorization task[52]. This study is consistent with ours in suggesting a benefit of overnight sleep for both unique and shared structure. They did not find any reliable above-baseline improvements, but there was numerical improvement in novel exemplar accuracy similar in magnitude to our shared feature effects, and statistical power may have been lower due to an across-subject design and fewer subjects.

Another study found increased generalization in an object categorization task over the course of an afternoon delay in both nap and wake conditions, with no difference between the two[53]. The task assessed memory or inference for the locations of faces, where some locations were predicted by feature rules (e.g. faces at one location were all young, stout, and had no headwear) and other locations had no rules. Overall memory for studied faces decreased over time, but more so for faces at locations without feature rules, suggesting a benefit due to shared structure in the rule location. While there are many differences between this paradigm and ours, our findings suggest the possibility that the forgetting and lack of sleep-wake differences observed may be due to averaging across all memories, instead of focusing on the weaker memories; Our Experiment 2 findings averaged across frequency are qualitatively similar to the findings from this study.

How does semantic memory consolidation work in the brain? The current results are consistent with the idea that initial memories of object categories are hippocampally dependent but that sleep serves to increase reliance on cortical representations. As cortical representations are more overlapping, they are better suited to representing shared structure, allowing an above-baseline benefit not seen for object-unique features. Because our task involves learning across visual and verbal modalities, we would expect the anterior temporal lobe to be the site of such cortical consolidation[54]. Another possibility to consider is that consolidation occurring within the hippocampus itself is responsible for the improvement in shared feature memory. There is certainly hippocampal plasticity during sleep[55], and the CA1 subfield of the hippocampus, implicated in consolidation[56], may be capable of representing coherent structure using a moderately overlapping neural code[8]. Lastly, as noted above, the lack of sleep benefits on generalization leaves open another possibility, whereby shared feature memory is improved through strengthened hippocampal item traces interacting with overlapping representations stored (but not locally strengthened) in cortex. Future research will be needed to disentangle these possibilities.

Our findings also contribute to our understanding of semantic memory consolidation mechanisms in suggesting that sleep first prioritizes objects that are weakly learned and most in need of further processing, preventing forgetting of unique features and promoting knowledge of shared features for such items. Longer sleep leads to maintenance of unique features and enhancement of shared category features regardless of initial strength of learning. Overall, our results suggest that sleep actively shapes learning of semantic category structure while simultaneously preserving knowledge of individuating details.

References

1. Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn Sci* **10**, 309–318, https://doi.org/10.1016/j.tics.2006.05.009 (2006).
2. Rogers, T. T. & McClelland, J. L. *Semantic Cognition: A Parallel Distributed Processing Approach*. (MIT Press, 2004).
3. Keil, F. C. The MIT *Press series in learning, development, and conceptual change. Concepts, kinds, and cognitive development*. (The MIT Press, 1989).
4. Murphy, G. L. & Medin, D. L. The role of theories in conceptual coherence. *Psychol Rev* **92**, 289–316 (1985).
5. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev* **102**, 419–457 (1995).
6. Norman, K. A. & O'Reilly, R. C. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol Rev* **110**, 611–646, https://doi.org/10.1037/0033-295X.110.4.611 (2003).
7. Kumaran, D. & McClelland, J. L. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol Rev* **119**, 573–616, https://doi.org/10.1037/a0028681 (2012).
8. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **372**, https://doi.org/10.1098/rstb.2016.0049 (2017).
9. Rasch, B. & Born, J. About sleep's role in memory. *Physiological reviews* **93**, 681–766, https://doi.org/10.1152/physrev.00032.2012 (2013).
10. Landmann, N. *et al*. The reorganisation of memory during sleep. *Sleep medicine reviews* **18**, 531–541, https://doi.org/10.1016/j.smrv.2014.03.005 (2014).
11. Loftus, G. R. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **11**, 397–406 (1985).
12. Drosopoulos, S., Schulze, C., Fischer, S. & Born, J. Sleep's function in the spontaneous recovery and consolidation of memories. *J Exp Psychol Gen* **136**, 169–183, https://doi.org/10.1037/0096-3445.136.2.169 (2007).
13. Peters, K. R., Smith, V. & Smith, C. T. Changes in sleep architecture following motor learning depend on initial skill level. *J Cogn Neurosci* **19**, 817–829, https://doi.org/10.1162/jocn.2007.19.5.817 (2007).
14. Diekelmann, S., Born, J. & Wagner, U. Sleep enhances false memories depending on general memory performance. *Behav Brain Res* **208**, 425–429, https://doi.org/10.1016/j.bbr.2009.12.021 (2010).
15. Cairney, S. A., Lindsay, S., Sobczak, J. M., Paller, K. A. & Gaskell, M. G. The Benefits of Targeted Memory Reactivation for Consolidation in Sleep are Contingent on Memory Accuracy and Direct Cue-Memory Associations. *Sleep* **39**, 1139–1150, https://doi.org/10.5665/sleep.5772 (2016).
16. McDevitt, E. A., Duggan, K. A. & Mednick, S. C. REM sleep rescues learning from interference. *Neurobiology of learning and memory* **122**, 51–62, https://doi.org/10.1016/j.nlm.2014.11.015 (2015).
17. Kuriyama, K., Stickgold, R. & Walker, M. P. Sleep-dependent learning and motor-skill complexity. *Learning & memory* **11**, 705–713, https://doi.org/10.1101/lm.76304 (2004).
18. Djonlagic, I. *et al*. Sleep enhances category learning. *Learning & memory* **16**, 751–755, https://doi.org/10.1101/lm.1634509 (2009).
19. Sio, U. N., Monaghan, P. & Ormerod, T. Sleep on it, but only if it is difficult: effects of sleep on problem solving. *Mem Cognit* **41**, 159–166, https://doi.org/10.3758/s13421-012-0256-7 (2013).
20. Tucker, M. A. & Fishbein, W. Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep* **31**, 197–203 (2008).
21. Talamini, L. M., Nieuwenhuis, I. L., Takashima, A. & Jensen, O. Sleep directly following learning benefits consolidation of spatial associative memory. *Learning & memory* **15**, 233–237, https://doi.org/10.1101/lm.771608 (2008).
22. Chin-Parker, S. & Ross, B. H. The effect of category learning on sensitivity to within-category correlations. *Mem Cognit* **30**, 353–362 (2002).
23. Johns, M. W. Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep* **15**, 376–381 (1992).
24. Adan, A. & Almirall, H. Horne & Ostberg morningness-eveningness questionnaire: a reduced scale. *Personality and Individual Differences* **12**, 241–253 (1991).
25. Akerstedt, T. & Gillberg, M. Subjective and objective sleepiness in the active individual. *Int J Neurosci* **52**, 29–37 (1990).
26. Rechtschaffen, A. & Kales, A. *A Manual of Standardized Terminology, Techniques, and Scoring Systems for Sleep Stages of Human Subjects*. (Brain Information/Brain Research Institute UCLA, 1968).
27. Wamsley, E. J. *et al*. Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation? *Biol Psychiatry* **71**, 154–161, https://doi.org/10.1016/j.biopsych.2011.08.008 (2012).
28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* **57**, 289–300 (1995).
29. Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S. & Wixted, J. T. An opportunistic theory of cellular and systems consolidation. *Trends Neurosci* **34**, 504–514, doi:S0166-2236(11)00091-9 (2011).
30. Schreiner, T. & Rasch, B. To gain or not to gain - The complex role of sleep for memory: Comment on Dumay (2016). *Cortex*, https://doi.org/10.1016/j.cortex.2016.06.011 (2016).
31. Dumay, N. Sleep not just protects memories against forgetting, it also makes them more accessible. *Cortex* **74**, 289–296, https://doi.org/10.1016/j.cortex.2015.06.007 (2016).
32. Ellenbogen, J. M., Payne, J. D. & Stickgold, R. The role of sleep in declarative memory consolidation: passive, permissive, active or none? *Current opinion in neurobiology* **16**, 716–722, https://doi.org/10.1016/j.conb.2006.10.006 (2006).
33. Durrant, S. J., Taylor, C., Cairney, S. & Lewis, P. A. Sleep-dependent consolidation of statistical learning. *Neuropsychologia* **49**, 1322–1331, https://doi.org/10.1016/j.neuropsychologia.2011.02.015 (2011).
34. Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D. & Walker, M. P. Human relational memory requires time and sleep. *Proc Natl Acad Sci USA* **104**, 7723–7728, https://doi.org/10.1073/pnas.0700094104 (2007).
35. Frost, R. L. & Monaghan, P. Sleep-Driven Computations in Speech Processing. *PloS one* **12**, e0169538, https://doi.org/10.1371/journal.pone.0169538 (2017).
36. Fenn, K. M., Nusbaum, H. C. & Margoliash, D. Consolidation during sleep of perceptual learning of spoken language. *Nature* **425**, 614–616, https://doi.org/10.1038/nature01951 (2003).
37. Mednick, S., Nakayama, K. & Stickgold, R. Sleep-dependent learning: a nap is as good as a night. *Nat Neurosci* **6**, 697–698, https://doi.org/10.1038/nn1078 (2003).
38. Lau, H., Tucker, M. A. & Fishbein, W. Daytime napping: Effects on human direct associative and relational memory. *Neurobiology of learning and memory* **93**, 554–560, https://doi.org/10.1016/j.nlm.2010.02.003 (2010).
39. Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C. & Mednick, S. C. REM, not incubation, improves creativity by priming associative networks. *Proc Natl Acad Sci USA* **106**, 10130–10134, doi:0900271106 (2009).
40. Batterink, L. J., Oudiette, D., Reber, P. J. & Paller, K. A. Sleep facilitates learning a new linguistic rule. *Neuropsychologia* **65**, 169–179, https://doi.org/10.1016/j.neuropsychologia.2014.10.024 (2014).
41. Lerner, I. *et al*. The influence of sleep on emotional and cognitive processing is primarily trait- (but not state-) dependent. *Neurobiology of learning and memory* **134**(Pt B), 275–286, https://doi.org/10.1016/j.nlm.2016.07.032 (2016).
42. Diekelmann, S. & Born, J. The memory function of sleep. *Nat Rev Neurosci* **11**, 114–126, doi:nrn2762 (2010).

43. Nieuwenhuis, I. L., Folia, V., Forkstam, C., Jensen, O. & Petersson, K. M. Sleep promotes the extraction of grammatical rules. *PloS one* **8**, e65046, https://doi.org/10.1371/journal.pone.0065046 (2013).
44. Gomez, R. L., Bootzin, R. R. & Nadel, L. Naps promote abstraction in language-learning infants. *Psychol Sci* **17**, 670–674, https://doi.org/10.1111/j.1467-9280.2006.01764.x (2006).
45. Lau, H., Alger, S. E. & Fishbein, W. Relational memory: a daytime nap facilitates the abstraction of general concepts. *PloS one* **6**, e27139, https://doi.org/10.1371/journal.pone.0027139 (2011).
46. Pace-Schott, E. F. *et al*. Sleep promotes generalization of extinction of conditioned fear. *Sleep* **32**, 19–26 (2009).
47. Hennies, N., Lewis, P. A., Durrant, S. J., Cousins, J. N. & Ralph, M. A. Time- but not sleep-dependent consolidation promotes the emergence of cross-modal conceptual representations. *Neuropsychologia* **63**, 116–123, https://doi.org/10.1016/j.neuropsychologia.2014.08.021 (2014).
48. Werchan, D. M. & Gomez, R. L. Wakefulness (not sleep) promotes generalization of word learning in 2.5-year-old children. *Child Dev* **85**, 429–436, https://doi.org/10.1111/cdev.12149 (2014).
49. Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J. & Rauss, K. Sleep Supports the Slow Abstraction of Gist from Visual Perceptual Memories. *Scientific reports* **7**, 42950, https://doi.org/10.1038/srep42950 (2017).
50. Mirkovic, J. & Gaskell, M. G. Does Sleep Improve Your Grammar? Preferential Consolidation of Arbitrary Components of New Linguistic Knowledge. *PloS one* **11**, e0152489, https://doi.org/10.1371/journal.pone.0152489 (2016).
51. Tamminen, J., Lambon Ralph, M. A. & Lewis, P. A. The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *J Neurosci* **33**, 15376–15381, https://doi.org/10.1523/JNEUROSCI.5093-12.2013 (2013).
52. Graveline, Y. M. & Wamsley, E. J. The impact of sleep on novel concept learning. *Neurobiology of learning and memory* **141**, 19–26, https://doi.org/10.1016/j.nlm.2017.03.008 (2017).
53. Sweegers, C. C. & Talamini, L. M. 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 (2014).
54. Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat Rev Neurosci* **18**, 42–55, https://doi.org/10.1038/nrn.2016.150 (2017).
55. Grosmark, A. D., Mizuseki, K., Pastalkova, E., Diba, K. & Buzsaki, G. REM sleep reorganizes hippocampal excitability. *Neuron* **75**, 1001–1007, https://doi.org/10.1016/j.neuron.2012.08.015 (2012).
56. Remondes, M. & Schuman, E. M. Role for a cortical input to hippocampal area CA1 in the consolidation of a long-term memory. *Nature* **431**, 699–703, https://doi.org/10.1038/nature02965 (2004).

Acknowledgements

We thank Roy Cox and Robert Stickgold for helpful discussions and Christopher Cox, Nicholas Reihanabad, and Chalani Perera for help running subjects. This work was supported by: NIH NINDS F32-NS093901 (ACS); NSF GRFP (EAM); NIH NIA R01-AG046646 (SCM); NSF BCS-1439210 (SCM); NIH NIMH R01-MH069456 (KAN).

Author Contributions

A.C.S., K.A.N., and T.T.R. designed the study. L.C. managed data collection for Experiment 1. E.A.M. managed data collection and analyzed polysomnography data for Experiment 2. A.C.S. and T.T.R. analyzed behavioral data from both experiments. All authors discussed the interpretation of the data. A.C.S. drafted the manuscript and all authors edited the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12884-5.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.