Journal of Memory and Language 138 (2024) 104536

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Recall and recognition of discourse memory across sleep and wake

Matthew H.C. Mak [a,b,*], Adam J. Curtis [b], Jennifer M. Rodd [c], M. Gareth Gaskell [b,*]

[a] Department of Psychology, University of Warwick, UK
[b] Department of Psychology, University of York, UK
[c] Department of Experimental Psychology, University College London, UK

ARTICLE INFO

Keywords:
Discourse memory
Sleep
Distortion
Consolidation
Recall
Recognition

ABSTRACT

The episodic context account (Gaskell et al., 2019) proposes that the act of language comprehension gives rise to an episodic discourse representation, and that this representation is prone to sleep-related memory effects. In three experiments, we tested this prediction by asking participants to read/listen to naturalistic stories before their memory was tested after a 12-hr interval, which included either daytime wakefulness or overnight sleep. To assess discourse memory, we used sentence recognition (Experiment 1; N = 386), free story recall (Experiment 2; N = 96), and cued recall (Experiments 2 and 3; N = 192). We found no evidence of sleep-related effects in sentence recognition or free recall, but cued recall (aka fill-in-the-blank) showed that the degree of time-related distortion, as indexed by both a subjective categorisation measure and Latent Semantic Analysis, was lower after sleep than after wake. Overall, our experiments suggest that the effect of sleep on discourse memory is modest but observable and may [1] be constrained by the retrieval processes (recollection vs. familiarity & associative vs. item), [2] lie on a qualitative level that is difficult to detect in an all-or-nothing scoring metric, and [3] primarily situated in the textbase level of the tripartite model of discourse processing.

Introduction

A burgeoning body of evidence suggests that sleep influences language learning. Studies of infants (Friedrich et al., 2017; Horváth et al., 2015), children (James et al., 2020; Williams & Horst, 2014) and adults (Bakker, et al., 2014; Dumay & Gaskell, 2007; Wang et al., 2017) have converged to show that a period of post-exposure sleep (vs. an equivalent amount of wakefulness) often benefits the retention of newly acquired linguistic knowledge. On the neurocognitive level, this benefit is often attributed to sleep actively supporting memory consolidation (e.g., Born & Wilhelm, 2012; Davis & Gaskell, 2009; McClelland et al., 1995). These theories posit that the hippocampus steps in to enable rapid encoding of new linguistic knowledge, which might be replayed within the hippocampus during sleep-related consolidation and be progressively fed into long-term neocortical stores (although cf. Yonelinas et al., 2019 for an alternative characterisation).

Studies that have revealed clear sleep-related effects in the language domain tended to use novel linguistic materials; for instance, some word-learning studies trained participants on pseudowords such as *cathedruke* and *feckton* (e.g., Dumay & Gaskell, 2007; Takashima et al., 2014; Tamminen & Gaskell, 2013; Wang et al., 2017). As such, it is possible that stimulus novelty is a main factor that underlies the benefits of sleep-related processes. However, a recent addition to the literature suggests that the effect of sleep on language processing may be broader than previously suggested and may extend to everyday language comprehension.

Building on Rodd et al.'s (2013, 2016) word-meaning priming paradigm, Gaskell et al. (2019) exposed participants to ambiguous words (e.g., *bark*) in sentences that biased interpretation towards the words' less common, subordinate meaning (e.g., *The branches and the bark had been damaged by the storm*). After 12 h including overnight sleep or daytime wakefulness, participants completed an associate production task, where they generated an associate for the target ambiguous words, presented in isolation. Here, participants in the sleep (vs. wake) group showed greater priming such that they generated significantly more associates related to the ambiguous words' subordinate meanings, suggesting that sleep-related memory effects are not restricted to novel linguistic materials. In a second experiment, Gaskell et al. (2019) tested whether these sleep-related effects were due to sleep actively stabilising the priming effects or simply protecting them from wakeful interference.

* Corresponding authors at: Department of Psychology, University of Warwick, Coventry CV4 7AL, UK (M.H.C. Mak); and Department of Psychology, University of York, Heslington, York YO10 5DD, UK (M.G. Gaskell).
E-mail addresses: matthew.mak@warwick.ac.uk (M.H.C. Mak), gareth.gaskell@york.ac.uk (M.G. Gaskell).

https://doi.org/10.1016/j.jml.2024.104536
Received 21 February 2023; Received in revised form 19 April 2024; Accepted 20 May 2024
Available online 30 May 2024

To do so, the original 12-hour delay between exposure and association production was extended to 24 h, during which half of the participants had a period of overnight sleep before daytime wakefulness (Sleep-Wake) while the other half did the opposite (Wake-Sleep). Gaskell et al. found that participants in the Sleep-Wake group showed stronger priming, suggesting that sleep made priming resistant to interference during the following day awake, thereby providing evidence for sleep having an active role to play in maintaining the priming effect. These findings were interpreted with respect to an active consolidation account: When an ambiguous word is encountered, participants make use of the surrounding context to determine its intended interpretation. Gaskell et al. (2019) argued that in determining the intended meaning, some kind of associative learning is involved, resulting in an episodic representation that binds together the ambiguous word and its surrounding sentential context. This episodic representation, presumably supported by the hippocampus (Milivojevic et al., 2016), is subject to decay, but if a sleep opportunity follows shortly after sentence exposure, it might be more likely to be stabilised and/or strengthened by sleep-related consolidation. As a result, this representation might be able to bias participants in associate production to generate a response that is consistent with the ambiguous words' subordinate meaning. The finding that sleep-related effects extended to the comprehension of highly familiar lexical units led Gaskell et al. (2019) to propose an episodic context account (originally called contextual binding account), which postulates a broader role of sleep in day-to-day language comprehension.

Consider the episodic representation once again. Presumably, it binds together the elements in the sentential context in which an ambiguous word occurs (Gaskell et al., 2019). By extension, encounters with all kinds of meaningful utterances or texts—regardless of whether they contain ambiguous words—may result in a hippocampus-dependent episodic representation. Note that this representation needs not be a carbon copy of the sentence; instead, it may be relatively abstract and contain its gist (Curtis et al., 2022; Gaskell et al., 2019). If language comprehension indeed gives rise to this kind of episodic representation, it would mean that memories for all kinds of sentences could be influenced—perhaps to a varying degree—by sleep-related processes (Gaskell et al., 2019; Mak, Curtis, et al., 2023). Motivated by this possibility, we ask in this paper whether memory representations derived from naturalistic stories are influenced by sleep (vs. wakefulness).

Drawing from both the episodic memory and sleep literatures, the episodic context account (Gaskell et al., 2019) posits that hippocampus-dependent episodic memory is routinely involved in language comprehension (see also Blank et al., 2016; Duff & Brown-Schmidt, 2012, 2017). Specifically, it proposed that at the point of comprehension, episodic memory would step in to enable the rapid binding of various discourse elements (e.g., characters, spatio-temporal contexts), forming a context-specific representation that can be used to guide on-line comprehension and/or facilitate the construction of an event model (e.g., Kintsch, 1994). Importantly, given the episodic nature of this discourse representation, its retention is predicted to be prone to sleep-related memory effects. Notably, this prediction is not specific to the episodic context account, as general accounts of memory/sleep predict that sleep would benefit the consolidation of declarative memory, regardless of the specific type of material (e.g., Rasch & Born, 2007). The distinguishing factor of the episodic context account lies in its core as a theory of language comprehension, which emphasises the involvement of episodic memory and sleep in language comprehension. As far as we can see, this is a first in the literature as existing theories on language comprehension (e.g., Kintsch, 1992; Kintsch, 1994) typically do not consider the effects of sleep; for instance, recent neurocognitive models of language comprehension (e.g., Blank et al., 2016; Duff & Brown-Schmidt, 2012, 2017) suggest that episodic memory networks may contribute to some aspects of language comprehension, but these theories do not explicitly address the influence of sleep. In fact, most models of comprehension would at best remain agnostic about how sleep may influence comprehension (e.g., Duff, et al., 2020). Therefore, our research aims to bridge the literature on language comprehension and sleep by marrying these two domains. In other words, we are pushing theoretical boundaries.

As mentioned above, the prediction that sleep may influence memory for discourse is not specific to the episodic context account, so it is not surprising that a few prior studies in the memory literature have tested this prediction. Unfortunately, however, the existing evidence base is small and somewhat inconsistent. Below, we briefly review these studies.

Sleep and discourse memory

Although not primarily interested in discourse memory *per se*, Wagner et al. (2001) was among the first to use naturalistic stories in a sleep study, where 23 young adults read four short stories, two on an emotional topic (e.g., a murder) and two on a neutral topic (e.g., bronze making). Each story was recalled on two occasions, once immediately after exposure and once after a 3-hour delay filled with either sleep or wakefulness. Discourse memory was quantified via the number of content words recalled, defined as the verbatim words used in the stories, close synonyms to the verbatim words, as well as word type transitions (e.g., from noun to adjective). The study reported that participants who had a sleep opportunity (vs. those who stayed awake) retained more content words from the emotional stories (The Cohen's *d* for this sleep benefit is estimated to be 2.93), although this between-group difference did not hold in the neutral stories. It is thus concluded that whether sleep benefits discourse memory is dependent on the emotionality of the content (see also Reid et al., 2022).

In contrast to Wagner et al. (2001), Aly and Moscovitch (2010) found a clear sleep benefit in the recall of neutral stories. Instead of counting content words, Aly and Moscovitch (2010) scored a participant's recollection based on the number of story propositions (i.e., ideas irrespective of wording). They found that both younger ($N = 10$) and older ($N = 12$) adults retained more propositions if recall was preceded by sleep than by wakefulness. Note that the reported effect size for this sleep benefit was large, at Cohen's $ds > 1$; this means that in order to achieve >80 % statistical power, a total sample size of >32 participants in a between-participant design is required (assuming alpha = 0.05 and *t*-test being used). In a more recent study, Cohn-Sheehy et al., (2022; Experiment 2) tested the effect of sleep on free story recall in a total of 90 young adults. Their study made use of neutral stories that incorporated main and side plots, as well as story elements that were either coherent or incoherent. As in Aly and Moscovitch (2010), all participants in Cohn-Sheehy et al. had an immediate retrieval practice before a period of overnight sleep or daytime wakefulness. Despite being well-powered and using a scoring metric akin to Aly and Moscovitch's, Cohn-Sheehy et al. found no evidence of any sleep-dependent benefits across all their story manipulations, and the effect size of sleep was negligible, estimated to be in the region of Cohen's $d = 0.04$–0.1. In a study with a similar sample size ($N = 94$) and experimental design, van Rijn et al. (2017) also found that sleep did not exert a significant effect on discourse memory, and the effect size of sleep was estimated to be Cohen's $d < 0.1$. Furthermore, Schönauer et al. (2014) tested how sleep (vs. wake) may influence declarative memories for a range of learning materials, one of which being neutral stories. In three experiments (each with a sample size of ~17), there was no evidence that sleeping soon after story encoding benefitted story recall. Interestingly, when the results from these experiments were combined and analysed as one, a sleep benefit emerged. However, caution is warranted in data interpretation given the exploratory nature of their pooled analysis.[1]

[1] Note that Schönauer et al. (2014) did not provide the effect size of sleep or report enough information for us to estimate it.

One possible explanation for why some studies did not find a clear effect of sleep on discourse memory is that its benefit might be occluded by repeated testing, as employed by both Cohn-Sheehy et al. (2022) and van Rijn et al. (2017). Bäuml et al., (2014; Experiment 4B) demonstrated a substantial sleep benefit in discourse memory (Cohen's $d = 1.13$) when participants did not retrieve the stories immediately after encoding ($N = 24$). However, this sleep advantage was not evident when participants ($N = 24$) were tested on the story both before and after the delay interval (Cohen's $d = 0.2$). This finding suggests that repeated testing can potentially reduce or even eliminate the benefit of sleep on discourse memory (see also Abel et al., 2019)—a possibility that we attempted to address in free story recall in Experiment 2. In sum, the findings from the above 'sleep vs. wake' studies are inconsistent, and this is intriguing, especially in light of the consistent benefit that sleep confers on discrete and static stimuli like paired associates (e.g., Plihal & Born, 1997; Lo et al., 2014; Scullin, 2013).

Finally, a different yet related line of research, which also yielded varied findings, stems from studies on sleep deprivation. In Tilley and Empson (1978), 20 participants were tested for retention of a story by means of free recall after being woken up for a few minutes at the onset of either Rapid Eye Movement (REM) sleep or Stage-4 sleep. Those woken at the onset of REM sleep showed poorer story recall, suggesting that this sleep stage may be particularly important for memory consolidation (see also Empson et al., 1981). In contrast, a study by Blagrove and Akehurst (2000) revealed no effect of total sleep deprivation on story retention: A total of 93 participants, who were either sleep deprived for 29–50 h or had normal sleep-wake cycles during those hours, were tested on story recall. Interestingly, the two groups performed similarly, questioning the extent to which sleep is involved in the retention of discourse memory.

Here, we speculate why discourse memories, especially those derived from stories with a neutral topic, might not be consistently affected by sleep (vs. wake). First, some existing evidence suggests that sleep confers a larger benefit on weakly (vs. strongly) encoded declarative memories (Denis, Mylonas, et al., 2021; Schoch et al., 2017). Compared to discrete and static stimuli like paired associates, elements in a naturalistic discourse tend to be more strongly related, due to, for example, the presence of causal links (Radvansky, 2012). This implies that elements in naturalistic stories are generally encoded with greater strength than those in discrete and static stimuli. In turn, this may reduce any potential benefit that sleep may bring to discourse memory (Cohn-Sheehy et al., 2022). Second, in tapping discourse memory, almost all prior studies used free recall and an all-or-nothing scoring approach, meaning that one point was awarded for every correctly recalled content word or proposition (e.g., Aly & Moscovitch, 2010; van Rijn et al., 2017). An implication, then, is that recalled elements not mentioned in the stories, such as inferences and errors, were discarded from analysis. This is not an issue *per se*, but the effect of sleep on discourse memory may be more nuanced than what can be captured in an all-or-nothing manner. Findings from a recent study support this view.

Denis, Dipierto, et al. (2021) exposed American undergraduates to a Native American folklore, entitled *War of the Ghosts*. First used by Bartlett (1932), the story was written in a non-Western style that is unfamiliar to most, if not all, people in Western countries. This implies relatively weak associative links between discourse elements in these stories, and hence, they are likely to be encoded with weaker strength compared to what would be expected for a typical Western story. Following a 12-hr delay filled with either overnight sleep or daytime wakefulness, participants recalled the story in a free recall procedure, which was scored by assigning each recalled proposition into one of seven categories: accuracy, gist, omission, inference, normalisation, incorrect placement, and importation. Denis, Dipierto, et al. (2021) found that the Sleep group outperformed the Wake group in terms of accuracy (i.e., more veridical propositions being recalled). Potentially, this is related to *War of the Ghosts* having weakly associated discourse elements, which, in turn, increased the likelihood of sleep exerting an effect. Furthermore, Denis et al. also found that the Sleep (vs. Wake) group recalled more events that were not explicitly stated but could be reasonably inferred from the story (inference), and relatedly, these participants also recalled fewer fabricated events (importations). In other words, less distortion seemed to have taken place in those who had a sleep opportunity, providing support for the proposal that a more nuanced scoring approach might be needed to fully capture the effect of sleep on discourse memory. Despite this, it is unclear if the findings from Denis, Dipierto, et al. (2021) generalise to schema-consistent stories, as the War of the Ghosts story represents the extreme end of the schema-consistency spectrum, which is rare in daily life. In order to further our understanding of the interplay between sleep and discourse memory, it is important to increase ecological validity by using stories that are more comprehensible and representative of natural language (e.g., stories with discourse elements that have stronger associative links). In addition, the scoring system adopted by Denis, Dipierto, et al. (2021) is subjective in nature; for instance, the distinction between inference and importation is debatable—what is considered an inference by one person may be considered an importation by another. This makes it hard for future studies to replicate and highlights a clear need for more objective scoring metrics (a goal we aim to achieve in the current research).

In sum, the existing evidence base concerned with the effect of sleep on discourse memory is not only inconsistent, but it is also dominated by evidence built upon an all-or-nothing scoring metric that may not fully capture qualitative changes to discourse memory. In light of this, we conducted three Experiments, designed to test how memories for naturalistic stories might be influenced by sleep (vs. wakefulness). In tapping discourse memory, we made use of a well-established sentence recognition procedure (Experiment 1) and a novel cued recall task (Experiments 2 and 3), both of which allowed us to assess discourse memory in a nuanced manner. We also replicated Aly and Moscovitch (2010) in Experiment 2 by quantifying discourse memory via free recall. These experiments represent arguably the most comprehensive examination to date of the effect of sleep on discourse memories, providing us with an opportunity to test a key prediction of the episodic context account and to reconcile the existing literature. To help readers better understand our experimental procedures and their relation to the overarching question, we summarised our series of experiments in Table 1.

Experiment 1

Experiment 1 is built upon a tripartite model espoused by decades of research in discourse comprehension (e.g., Fletcher & Chrysler, 1990; Kintsch & van Dijk, 1978; Kintsch et al., 1990; Seger et al., 2021; van Dijk et al., 1983; Zwaan & Brown, 1996). According to this model, comprehenders form three different, although interrelated, mental representations of verbal text: *surface*, *textbase*, and *event model*. The surface level refers to memory for the exact wording, whereas textbase is concerned with the propositions, regardless of the wording. A sentence

Table 1
Summary of the experimental procedures and research questions.

Exp	Positive control task	Main task	Overarching question of main task	Specific question of main task
1A	Free wordlist recall	Sentence recognition	Does sleep influence memory for discourse?	Does sleep differentially affect the three levels (surface, textbase, event model) of discourse memory?
1B	Paired-associate learning			
2	Free story recall	Fill-in-the-blank		Does sleep influence the distortion of discourse memory?
3	Paired-associate learning			

such as "Peter was starving" would be identical to "Peter was hungry" on the textbase level but different on the surface form. Then, for the event model, text information is elaborated with reference to the comprehender's prior knowledge, producing inferences; for example, in reading "Peter was starving", one may infer that Peter has not eaten for a while or that his stomach is growling audibly.

In the discourse processing literature (Fletcher & Chrysler, 1990; Kintsch et al., 1990; Zwaan, 1994), the three levels of discourse representation are typically indexed via a sentence recognition paradigm, pioneered by Schmalhofer and Glavanov (1986). In this paradigm, participants read a narrative story, followed by a recognition task, where participants judge if a particular sentence was previously read in the story. Four types of probe are presented: *verbatim*, *paraphrase*, *inference*, and *wrong* (see Table 2 for details). With these probes, it is possible to derive memory estimates of the three levels of discourse representation using signal detection measures.

The measure of surface memory compares judgements for the verbatim probes (Hits) with judgements for the paraphrases (False alarms). Both probe types convey propositions mentioned in the story, but only the verbatim probes contain the actual wording and syntax. If participants have excellent surface memory for the sentences, then they will be able to accurately discriminate between these two probe types (i.e., accept verbatim probes, while rejecting paraphrases). In contrast, if participants' memory for the sentence has lost this level of surface detail, then they will be performing close to or at chance level at this discrimination. Next, the measure of textbase memory compares judgements for the paraphrases (Hits) with judgements for the inferences (False alarms). These probe types are consistent with the event described in the text, but only the paraphrase conveys propositions that were actually present; therefore, participants who retained more textbase information should be more able to discriminate these (i.e., accept paraphrase, while rejecting inferences). Finally, the event model measure compares judgements for the inference probes (Hits) with judgements for the wrong probes (False alarms). Neither were mentioned, nor did they convey propositions present in the story; however, only the inferences were consistent with the event described in the story; therefore, readers with a high-quality event model should be better at discriminating these (i.e., accept inferences, while rejecting the wrong probes).

Using this sentence recognition paradigm and the signal detection measure described above, Fisher and Radvansky (2018) tracked the retention of the three levels of discourse representation in almost 300 participants, across delays of up to 12 weeks. They found that the three levels differed in their longevity: Surface information was forgotten soon after story exposure, although not completely. In contrast, retention of textbase memory and event model was substantially better: Textbase memories were well retained until about a week after story exposure, although memories for the event model were consistently higher throughout 12 weeks (see also Doolen & Radvansky, 2021). Fisher and Radvansky suggested that offline consolidation might play a role in the forgetting of the three levels over time; however, to the best of our knowledge, no existing studies have explicitly tested this possibility.

Table 2
Probe types in sentence recognition (Schmalhofer & Glavanov, 1986).

Probe types	Descriptions	Examples
Verbatim	A sentence that appeared in the story (exact wording and syntax).	*People were protesting against the war.*
Paraphrase	A sentence that did not actually appear in the story, although that same propositional idea was conveyed.	*People took part in an anti-war protest.*
Inference	A sentence that conveys an idea that was likely generated by readers using their world knowledge.	*People in the protest were angry.*
Wrong	Thematically consistent with the text, but it's inconsistent with the events described by the text.	*People in the protest supported the war.*

There are reasons to believe that sleep might differentially affect the three levels of discourse representation. For instance, active consolidation accounts argue that one of the key functions of sleep is to facilitate the integration of newly encoded memories into pre-existing knowledge (e.g., Davis & Gaskell, 2009). On this view, generation and retention of event models may be particularly prone to sleep-related effects, because such models require integrating the affairs described in a story with a comprehender's prior knowledge (Altmann & Ekves, 2019; Kintsch & van Dijk, 1978). However, it is also possible that the effect of sleep may primarily lie on the textbase level, because (1) surface information tends to be quickly forgotten (Sachs, 1974), leaving little room for sleep to exert an effect, and (2) event models tend to be fairly well retained after months and sometimes years (Doolen & Radvansky, 2021), leaving limited room for sleep to boost its retention. However, no prior "sleep × discourse" studies have attempted to tease apart the three levels of discourse representation—a key research gap that Experiment 1 aims to fill.

Design Overview

Experiment 1 was modelled upon Fisher and Radvansky (2018), comprising a study and a test phase, as summarised in Fig. 1.

Participants were randomly assigned to one of the four groups: Immediate-AM, Immediate-PM, Delay-Wake, and Delay-Sleep. The Immediate groups served as the control to rule out time-of-day effects, and they completed the test phase immediately after the study phase. Those in the AM group started the experiment at 9AM (±1 hr) while those in the PM group at 9PM (±1 hr). In the Delay groups, the test phase took place approximately 12 h after the study phase. Participants in the Wake group started the study phase at 9AM (±1 hr) and the test phase at 9PM (±1 hr) on the same day. Those in the Sleep group started the study phase at 9PM (±1 hr), and after 12 h including a period of overnight sleep, they carried out the test phase at 9AM (±1 hr) the following day. To sum up, the study had a 2 (Interval: Immediate vs. Delay) × 2 (Start Time: 9AM vs. 9PM) between-participant design.

Experiment 1 contains two sub-experiments, 1A and 1B, that differ only in the positive control task. Due to the COVID-19 pandemic, this study was conducted online via Prolific (https://www.prolific.co), so it was important to include a positive control test for which we would expect a sleep effect. In Experiment 1A, we used free wordlist recall (as in Fisher & Radvansky, 2018) as our positive control, which has previously been shown to benefit from a period of sleep (Lahl et al., 2008; Saletin et al., 2011; see also Abel & Bäuml, 2013; Drosopoulos et al., 2005 who used similar stimuli but different outcome measures). However, to foreshadow our results, no sleep-related benefit was observed in 192 participants, in contrast to our prediction and prior evidence. This motivated us to switch from free wordlist recall to paired-associate learning in Experiment 1B (N = 192). Replicating prior lab-based studies (e.g., Payne et al., 2012; Plihal & Born, 1997), this task revealed a clear and robust sleep benefit, giving us confidence that sleep-related memory effects can be detected both within our sample and in an online study (see also Ashton & Cairney, 2021). Regardless of the positive control task, participants in both Experiments 1A and 1B subsequently read four naturalistic stories, and their discourse memory was then indexed via the sentence recognition task described earlier. This means that a total of 384 participants completed sentence recognition. Both Experiments 1A and 1B, including the exclusion criteria and analysis plans, were pre-registered ahead of data collection (Experiment 1A: https://aspredicted.org/blind.php?x=j7xv5z/ and Experiment 1B: https://aspredicted.org/blind.php?x=t2bm6p). Any deviations are explicitly noted.

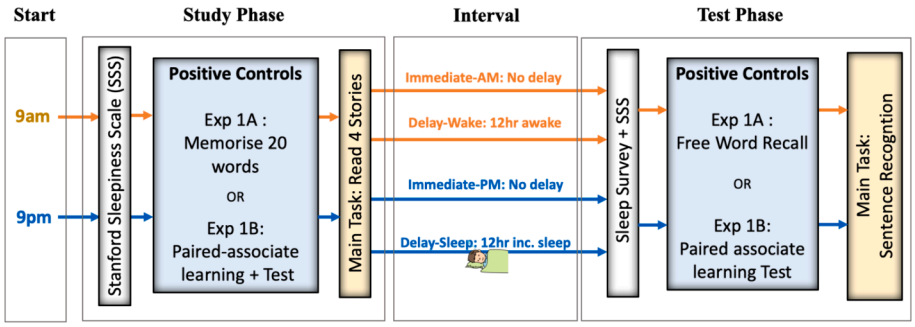


Fig. 1. Procedure of Experiment 1.

Methods

Participant recruitment

Following our prior studies (Mak, Curtis, et al., 2023; Mak, O'Hagan, et al., 2023; Ball et al., 2024), participants recruited from Prolific first filled out a screening survey, where they provided basic demographic information, read about the details of the main study, and indicated whether they wanted to take part in it ($N_{Exp1A} = 302$, $N_{Exp1B} = 329$). Our inclusion criteria were: (i) Aged between 18 and 25, (ii) English as (one of) their first language(s), (iii) Current resident in the UK, (iv) No known history of any psychiatric, developmental, or sleep disorders, and (v) Willing to be randomly allocated to one of the four groups.[2] We screened out respondents who did not meet these criteria, leaving us with 237 respondents in Experiment 1A and 272 in 1B. They were then randomly allocated to one of the four groups, with each receiving an invitation to take part in the main study at a specific time. We conducted the recruitment and screening process iteratively such that after the initial round of invitations and responses, we examined the number of participants in each group and identified any imbalance. Additional invitations were then sent to eligible individuals in the underrepresented group until a balanced distribution was reached. Of those who took up the invitation, 196 and 198 respondents completed all sessions in Experiment 1A and 1B respectively. Four were excluded from Experiment 1A and six from 1B for meeting our exclusion criteria of a sleepiness rating of six or above on the Stanford Sleepiness Scale ($N = 4$), reporting a nap between the study and test phases in the Wake group ($N = 2$) or less than six hours of sleep or poor sleep quality in the Sleep group ($N = 4$). Experiments 1A and B therefore each had usable data from 192 participants.

Participants

The combined sample size of Experiment 1A and B was 384 (261 females, 123 males; $M_{age} = 24.9$, $SD_{age} = 4.2$), evenly split between the four groups (i.e., 96 participants per group). This sample size was informed by Fisher and Radvansky (2018), who tracked time-related changes in discourse memory in groups of 48. We doubled the number as we needed sufficient power to detect an interaction between Interval (Immediate vs. Delay) and Start Time (9AM vs. 9PM) to infer any sleep-related effects.

[2] A limitation of our recruitment procedure, as pointed out by an anonymous reviewer, is that we may have introduced a selection bias into our sample. A portion of eligible participants declined the invitation, suggesting that these individuals may have had scheduling conflicts or preferences that did not align with the assigned time slots, resulting in a selection bias. Reassuringly though, as reported below, our wake and sleep groups were well-matched on their degree of sleepiness and morningness/eveningness preference. We believe our random group assignment is likely the most effective approach achievable in an online sleep study, offering a certain level of control over potential confounding variables (e.g., time-of-day preference).

Materials and procedure

Study phase

Wordlist (Positive control in 1A). We used nine different lists of 20 English words (e.g., *breathe, circle)*, taken from Mak et al. (2021) and Mak and Twitchell (2020).[3] Appendix A shows a sample wordlist and a summary of the lexical properties of these words. Participants were exposed to one of the nine wordlists in the study phase, randomly selected by the computer with equal probability. Each word was presented for 10 s each, whose order of presentation was randomised. At the beginning of the task, participants were informed that they would need to recall the words later on but that the words need not be recalled in the order they were shown.

Paired-associate learning (Positive control in 1B). Forty cue-target pairs were taken from a previous study (Ashton & Cairney, 2021; Experiment 1). Words in each pair (e.g., *trophy – prize*) were semantically related, although the associative strength was relatively low ($M = 0.132$; $SD = 0.129$; Nelson et al., 2004). Appendix B shows the full set of word pairs used.

In the study phase, each cue-target pair was presented at the centre of the screen for 5000 ms. The next pair appeared after a 100 ms blank screen. Order of presentation was randomised. Participants were informed at the beginning that their memory for the word pairs would be subsequently tested. Four attention checks, where participants reported three digits (e.g., 531) immediately after they were displayed for 5000 ms, were included to ensure that participants paid sufficient attention (e.g., Mak, Curtis, et al., 2023; Mak, 2021). All participants passed at least three of them, so no data were discarded on this basis.

Immediately after exposure to the word pairs, all participants completed a baseline cued recall task. They were shown the cue from each pair and had 10 s to recall the target word by typing it out on their keyboard. The next trial began once a response was submitted (by hitting the return button) or if no response was recorded after 10 s. Regardless, participants received immediate feedback (a green tick or a red cross) and the correct answer following each trial, shown together on the screen for 1000 ms.

Story reading (Main task). Following Fisher and Radvansky (2018), we used the four naturalistic stories with a relatively neutral topic developed by Radvansky et al. (2001). These stories had an average of 621 words ($SD = 79$) and were entitled: Identification in the CIA, Farmer's Rebellion, New York in 2084, and Beanie Baby Craze (see Appendix C1 for Identification in the CIA).

In the study phase, the four stories were presented in a randomised order, with each preceded by its title, displayed alone on the screen for 5 s. The stories were then presented clause by clause in a self-paced

[3] Fisher and Radvansky (2018) also included wordlist recall as a positive control for their study. We followed their design by exposing participants to 20 words. However, they used a fixed list of 20 words, while we used nine lists of 20 words. This was intended to minimise stimulus-specific effects.

manner, with each clause replacing the last one. Participants advanced by clicking a "Next" button at the bottom of the screen. To reduce the possibility that participants would click the Next button without actually reading, the button only appeared on screen 0.5 s after the clause was shown. Participants were instructed to read the stories carefully as their memory would be tested later on.

Test phase

The test phase began with a short survey, where participants gave a Stanford Sleepiness Scale (SSS) rating and completed the reduced version of the Morningness-Eveningness survey (Adan & Almirall, 1991; Horne & Östberg, 1976). Participants in the Delay-Wake group were asked to indicate whether they had napped in between the study and test phases, and if they did, how long the nap was. For the Delay-Sleep group, participants indicated the time they went to bed the night before and the time they woke up that morning. They also rated their sleep quality on a scale of 0 to 5, with 0 being very poor and 5 being excellent.

Free wordlist recall (Positive control in 1A). Participants had three minutes to type out as many of the 20 words as they could recall from the wordlist presented at study. Participants could not proceed until the time was up.

Paired-associate learning (Positive control in 1B). This is the same cued recall task as in the study phase, except participants did not receive feedback or the correct answer here.

Sentence recognition (Main task). Each story was associated with 64 sentence probes, taken from Fisher and Radvansky (2018). Each probe belonged to one of these categories: verbatim, paraphrase, inference, and wrong (see Table 2 and Appendix C2 for examples). Each probe was based on a different sentence from the story.

In the task, participants judged whether a probe sentence had been read earlier in the study phase. They were informed that the probes might be similar to the ones they had read but contained changes in wording. A decision was made by pressing the A key on their keyboard to indicate "Yes, I did read this sentence" or the L key to indicate "No, I had not read this sentence". A total of 256 probes were presented (64 probes × 4 stories), which were blocked by story. Block order was randomised, and so was the trial order within block. At the beginning of each block, participants were given the title of the story to indicate on which story the probes were based.

Results

Group characteristics

As summarised in Table 3, the four groups were highly comparable in terms of various key characteristics (e.g., level of sleepiness, morningness/eveningness). One-way ANOVAs comparing SSS and morningness/eveningness scores revealed no significant between-group differences (*ps* > .54).

Deviations from pre-registration

Experiments 1A and 1B were originally designed as two separate experiments, but analysing the sentence recognition data separately or as a combined dataset revealed essentially the same results, so we opted for the latter for increased power and simplicity's sake. We note that we peeked at the data upon completion of Experiment 1A, so we reduced the alpha level from 0.05 to 0.025 for sentence recognition to guard against Type-1 inflation error (Sagarin et al., 2014). All the analyses presented below followed our pre-registered analysis plans, although two analyses (i.e., one examining the likelihood with which a probe received a Yes judgement, another examining response bias as indexed by B" bias score) were relegated to the supplementary materials (available on Open Science Framework) to help make the main text more concise.

Table 3
Group characteristics of the four groups in Experiment 1. Values in parentheses indicate standard deviations.

Characteristics	Immediate-AM	Immediate-PM	Delay-Wake	Delay-Sleep
N of participants	96	96	96	96
N of females	62	66	68	65
Mean SSS at study (Max = 7)	2.70 (1.33)	2.47 (1.75)	2.45 (1.15)	2.68 (1.21)
Mean SSS at test (Max = 7)	2.76 (1.20)	2.67 (1.44)	2.59 (1.33)	2.61 (1.58)
Mean reduced Morningness/Eveningness score	13.34 (3.73)	12.93 (3.54)	13.66 (3.85)	13.17 (3.34)
Mean *N* of hrs between study and test	NA	NA	10 hr 31 min (1 hr 4 min)	10 hr 51 min (1 hr 28 min)
Mean *N* of sleep hrs between study and test	NA	NA	NA	7 hr 32 min (1 hr 9 min)

Notes. (1) SSS stands for Stanford Sleepiness Scale; it ranges from 1 to 6, with higher values indicating greater sleepiness. (2) Reduced Morningness/Eveningness score ranges from 5 to 25, with higher values indicating greater morningness preference.

Confirmatory analyses

Free wordlist recall (Positive control; 1A). One participant from the Delay-Sleep group was excluded as they submitted no response. This analysis is therefore based on 191 participants (see Fig. 2 for summary). The data were analysed in a 2 (Interval: Immediate vs. Delay) × 2 (Start Time: 9AM vs. 9PM) between-participant ANOVA. If sleep benefitted free wordlist recall, we expected an interaction effect, along with a higher recall rate in the Delay-Sleep than in the Delay-Wake group.

There was a main effect of Interval [$F(1, 187) = 11.67$, $p < .001$, $\eta^2 = 0.059$], with participants in the Immediate groups recalling more words ($M = 45$ %, $SD = 21$ %) than those in the Delay groups ($M = 34$ %, $SD = 22$ %). There was no effect of Start Time [$F(1, 187) = 1.98$, $p = .161$, $\eta^2 = 0.01$], and the critical interaction was not significant [$F(1, 187) = 0.82$, $p = .368$, $\eta^2 = 0.004$]. In short, we found no evidence that a period of overnight sleep (vs. daytime wakefulness) enhanced recall. This prompted us to switch to paired-associate learning (Plihal & Born, 1997) as a positive control in Experiment 1B.

Paired-associate learning (Positive control; 1B). Unlike free wordlist recall, paired-associate learning measured baseline performance immediately after exposure in the study phase, allowing us to control for pre-existing individual differences in encoding/recall.

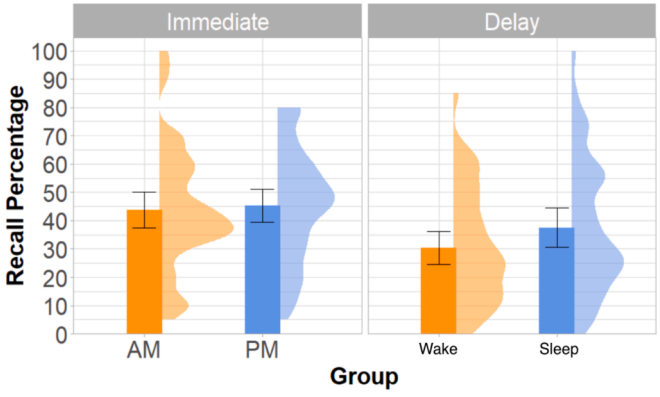


Fig. 2. Percentage of correct recall in free wordlist recall (Experiment 1A), summarised across groups. Note: Error bars represent 95 % between-subject CI while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

Performance at baseline and at Session 2 is summarised across groups in Table 4.

We first used a one-way ANOVA to check if baseline performance was comparable across the four groups, which revealed no significant difference [$F(3, 188) = 0.13$, $p = .941$, $\eta^2 = 0.002$].

Next, to test for a sleep effect, we performed a 2 (Interval: Immediate vs. Delay) × 2 (Start time: 9AM vs. 9PM) between-participant ANCOVA, with the number of correct recalls at test as the dependent variable and the number of correct recalls at baseline as the covariate. It revealed significant effects of Interval [$F(1, 187) = 46.93$, $p < .001$, $\eta^2 = 0.20$; Immediate > Delay], and Start Time [$F(1, 187) = 13.49$, $p < .001$, $\eta^2 = 0.07$, PM > AM]. Importantly, these were qualified by a significant interaction [$F(1, 187) = 5.15$, $p = .024$, $\eta^2 = 0.03$], which was followed up by two separate independent *t*-tests.

In these *t*-tests, the dependent variable was the difference in the number of correct recalls between the test phase and baseline (following Ashton & Cairney, 2021) (see Fig. 3). Within the Immediate condition, there was no significant difference between the AM and PM groups ($M_{AM} = 8.25$ vs. $M_{PM} = 9.21$), $t(83.02) = 1.10$, $p = 0.27$, $d = 0.22$. In the Delay condition, however, participants in the Sleep group ($M = 6.44$) outperformed those in the Wake group ($M = 2.88$), $t(92.88) = 4.17$, $p < .001$, $d = 0.85$. Therefore, in this assessment of episodic declarative memory, we observed a sleep-associated benefit (e.g., Ashton & Cairney, 2021; Plihal & Born, 1997; Wang et al., 2017), showing that the experimental parameters of the current study were suitable for detecting sleep-related memory effects.

Sentence Recognition (Main task). The number of Yes responses to each probe type is summarised across groups in Fig. 4.

Using the sentence recognition response data (summarised in Fig. 4), we calculated A′ signal detection values (Snodgrass & Corwin, 1988)—the non-parametric equivalent to *d′* in signal detection—as estimates of the three levels of discourse memory for each participant, as in previous studies (e.g., Fisher & Radvansky, 2018): surface form (hit = verbatim; false alarms = paraphrase), textbase (hit = paraphrase; false alarms = inference), and event model (hit = inference; false alarms = wrong). Since the hit rates were always greater than the false alarms rates, the formula for A′ sensitivity is 0.5 + [(Hits − False alarms)(1 + Hits − False alarms)]/[4Hit(Hits − False alarms)], with higher A′ scores indicating greater accuracy (Range = 0–1) and a score of 0.5 indicating chance level (see Fig. 5 for summary across Interval and Start Time).

We analysed the A′ scores in a set of 2 (Interval: Immediate vs. Delay) × 2 (Start Time: 9AM vs. 9PM) ANOVAs, one for each of the three levels of discourse representation (see Table 5).

First, there was a main effect of Interval (Immediate vs. Delay) in surface [$F(1, 380) = 16.7$, $p < .001$] and textbase memory [$F(1, 380) = 20.6$, $p < .001$], with participants showing a significant decline in performance after a 12-hr delay. In contrast, the A′ scores for event model did not significantly differ between the Immediate and Delay conditions [$F(1, 380) = 0.17$, $p = .67$], suggesting that memory for event models was well maintained over 12 h, in line with Fisher and Radvansky (2018).

Start Time (9AM vs. 9PM) showed a main effect on the textbase level [$F(1, 380) = 7.49$, $p = .006$], such that participants who read the stories in the evening had lower A′ scores on this level. Potentially, this suggests that reading strategies varied across time of day, such that reading strategies tend to be less literal in the evening (e.g., Petros et al., 1990; Natale & Lorenzetti, 1997). We do not interpret this finding further as we did not have any *a priori* hypotheses regarding circadian differences.

Finally, if sleep influenced a certain level of discourse representation, there should be a significant interaction between Interval and Start Time; however, this was not the case at any of the three levels (*F*s < 1.86, *p*s > .16).

To facilitate interpretation of the null interactions, we performed exploratory Bayesian ANOVAs with default priors, following the procedure outlined in Wills et al. (2020). For almost all the interactions, the Bayes Factors (BF_{10}) were smaller than 0.3 (see Table 5), suggestive of moderate evidence for the null hypothesis (Lee & Wagenmakers, 2013). The only exception was the textbase A′ recognition score, where the interaction had a BF_{10} of 0.38 ± 3.77 %, indicative of anecdotal evidence for the null hypothesis. To sum up, this Bayesian analysis revealed that a period of overnight sleep (vs. daytime wakefulness) is unlikely to have an effect on the retention of the three levels of discourse representations, at least not when discourse memory is indexed by sentence recognition.

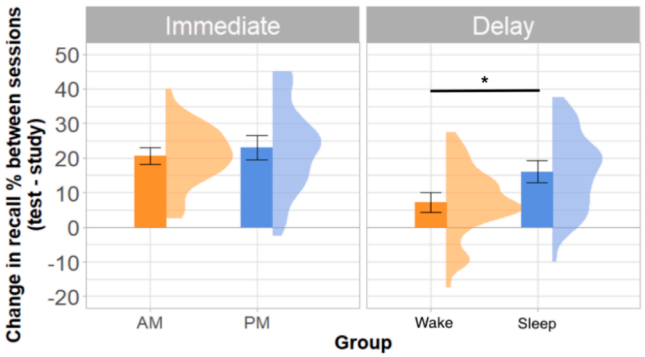


Fig. 3. Change in the number of correct recall between test and baseline in paired-associate learning (Experiment 1B), summarised across groups. Note. [1] * denotes statistical significance (p < .05). [2] Error bars represent 95 % between-subject CI while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

Table 4
Summary of cued recall performance in paired-associate learning (Experiment 1B).

Group	Mean % of correct recall at baseline (SD)	Mean % of correct recall in Session 2 (SD)	Changes across session
Immediate-AM	50.26 (18.9)	70.89 (18.9)	+20.63
Immediate-PM	47.81 (18.5)	70.83 (20.5)	+23.12
Delay-Wake	49.22 (20.5)	56.41 (21.0)	+7.19
Delay-Sleep	48.70 (20.2)	64.79 (19.4)	+16.09

Discussion

Experiment 1 made use of a well-established sentence recognition procedure to investigate how the retention of discourse memory is influenced by sleep (vs. wake). This procedure was designed to tease apart the three levels of discourse representation—surface, textbase, and event model, allowing us to go beyond what was explicitly mentioned in the texts and to examine the effect of sleep in a more comprehensive and nuanced manner. Contrary to the prediction of the episodic context account, we found no evidence that sleep was involved in the maintenance of any of the three levels. We are confident that this null result is *not* due to the study being conducted online, because one of our positive control tasks (i.e., paired-associate learning) showed a clear sleep-related benefit (see also Mak, Curtis, et al., 2023; Mak, O'Hagan, et al., 2023).

We begin by considering in turn the effect of sleep in the two positive control tasks. In free wordlist recall, we found no sleep-related effects, standing in contrast to prior studies that reported a large effect for sleep (Cohen's $d > 1$; Lahl et al., 2008; Saletin et al., 2011). An effect of this size means that our sample size of 48 participants/group gave us over 95 % statistical power to detect this sleep benefit. Our failure to replicate their findings suggest that the estimated effect sizes might be inflated due to relatively small sample sizes [e.g., Lahl et al., (2008; Experiment

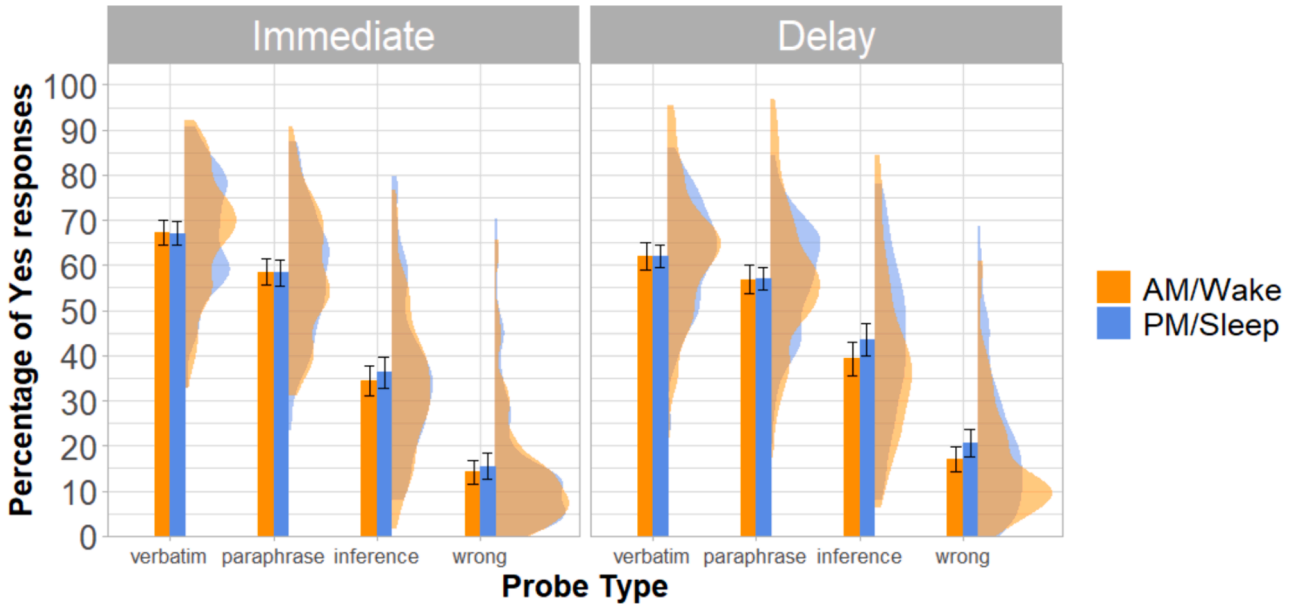


Fig. 4. Percentage of Yes responses to each probe type in sentence recognition (Experiment 1), summarised across groups. Note. Error bars represent 95 % between-subject CIs while the density functions represent the distribution of the data in each group. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

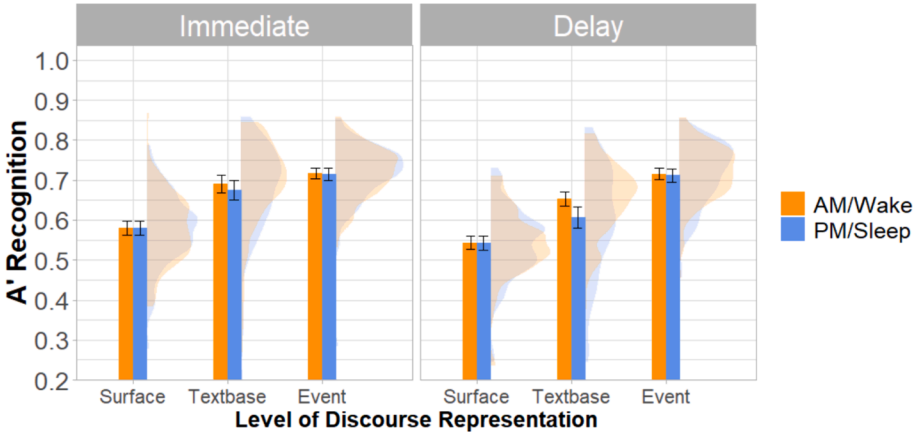


Fig. 5. A′ recognition scores across the three levels of discourse memory in Experiment 1, summarised across groups. Note. Error bars represent 95 % between-subject CIs while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

Table 5
ANOVA tables, alongside the Bayes Factors, for A′ recognition scores across the three levels of discourse representation in Experiments 1A + B.

A′ Recognition Score	Surface		Textbase		Event Model	
	ANOVA	BF_{10}	ANOVA	BF_{10}	ANOVA	BF_{10}
Interval (Immediate vs. Delay)	$F = 16.7$, $p < .001$*	326.83 ±0 %	$F = 20.6$, $p < .001$*	1582 ±0 %	$F = 0.17$, $p = .67$	0.12 ±0 %
Start Time (9AM vs. 9PM)	$F < 0.01$, $p = .99$	0.11 ±0 %	$F = 7.49$, $p = .006$*	3.40 ±0 %	$F = 0.15$, $p = .70$	0.12 ±0 %
Interval × Start Time	$F < 0.01$, $p = .99$	0.16 ± 2.61 %	$F = 1.86$, $p = .17$	0.38 ± 3.77 %	$F = 0.02$, $p = .89$	0.15 ± 2.44 %

Note. * denotes statistical significance ($p < .025$).

2) had 14 participants in a within-participant design, while Saletin et al. (2011) had 23 participants per group]. This possibility is supported by a recent registered report testing the effect of sleep (vs. wake) in the Deese-Roediger-McDermott recall paradigm (Mak, O'Hagan, et al., 2023), which had 240 participants, evenly split between a sleep and a wake group. Participants studied 120 words presented individually on a computer screen and recalled them 12 h later in free recall. While the sleep group recalled significantly more words, the effect size was small (Cohen's $d = 0.26$). To detect this in a one-sided t-test with 80 % power, over 180 participants per group is needed. Therefore, it is possible that the effect sizes reported by Lahl et al. and Saletin et al. were inflated due to their small sample sizes and that even with 48 participants/group, Experiment 1A was underpowered to detect a sleep benefit in free word recall. However, methodological factors may have contributed to the discrepant results between the current and prior studies. First, sleep was operationalised as a period of overnight sleep here but as a nap in both Lahl et al. (2008) and Saletin et al. (2011). Second, unlike Lahl et al. (2008) and Saletin et al. (2011), our participants read four stories after

studying the wordlist, which may have led to greater interference.

We then turn to paired-associate learning, which showed a clear sleep-related benefit. This gives assurance that within the precise sample that did sentence recognition, sleep-related effects were observable, even when the study was conducted online. As to why we saw a clear sleep benefit in paired-associate learning but not in free wordlist recall, there are multiple possible reasons. First, sleep may have a larger effect in tasks that tax associative memory (Diekelmann et al., 2009); we will revisit this point in General Discussion. Second, in paired-associate learning, we followed prior studies (e.g., Ashton & Cairney, 2021; Payne et al., 2012) by measuring retention at baseline, which was not the case in wordlist recall. Potentially, having a baseline measure may improve sensitivity to sleep-related benefits (Berres & Erdfelder, 2021; Lipinska et al., 2019). This prompted us to include baseline measures for all experimental tasks in the subsequent experiments.

Moving on, we consider the sentence recognition data. Putting the sleep manipulation aside first, our data are highly consistent with those from Fisher and Radvansky (2018): (i) surface memory was slightly above chance level in the Immediate groups, suggesting that surface memory was quickly, although not completely, forgotten soon after exposure, (ii) after a 12-hr delay, there was a significant decline in both surface and textbase memories, suggestive of time-related forgetting, and finally (iii) memory for the event model was high in both the Immediate and Delay groups, suggesting that it is relatively resistant to time-related forgetting.

Then, turning to the focus of our study—how the three levels of discourse representation may change over a period of overnight sleep (vs. daytime wakefulness). Contrary to the predictions of the episodic context account (or other general accounts of sleep), we found no evidence of any sleep-associated effects on the three levels of discourse representation. Furthermore, a Bayesian analysis revealed moderate evidence for the null hypothesis, suggesting that the three levels of discourse representation, as indexed by sentence recognition, were unlikely to have been affected by sleep. However, we cannot rule out the possibility that the null finding here is partially related to our Experiment 1 lacking a baseline measure for sentence recognition. Despite this possibility, the most parsimonious explanation of this null finding is that sleep has little or no role to play in discourse maintenance. This would argue against a strong version of the episodic context account and highlight a need for the theory to reconsider how central the role of sleep is in maintaining discourse memory. However, at this point, we have indexed discourse representation via only one outcome measure (i.e., sentence recognition). Before we draw any broader conclusion, we should consider whether other outcome measures, such as recall, might be useful to assess discourse memory.

There are reasons to believe that recall- and recognition-based procedures may differ in their sensitivity to sleep-related memory effects. In recall, participants are required to actively reconstruct a memory trace—a process that may not be required in recognition (Jacoby et al., 1993; Yonelinas, 1994, 2002; Yonelinas et al., 2010). On the neurocognitive level, recall procedures depend heavily on the hippocampus (e.g., Baddeley et al., 2001; Bastin et al., 2004; Girardeau et al., 2017; Holdstock et al., 2005; Mayes et al., 2002; Miyamoto et al., 2017; but see Squire et al., 2007), while recognition-based retrieval relies more on frontal-subcortical circuitry (Squire & Dede, 2015). As sleep is believed to be particularly important for the maintenance and/or consolidation of hippocampus-dependent memory, it has been argued that recall procedures are more likely to reveal sleep-related effects (Diekelmann et al., 2009). Motivated by this possibility, we used both free and cued recall in Experiment 2 to test for the effect of sleep on discourse memory.

Experiment 2

In keeping with the overarching aim, Experiment 2 tested whether memory for naturalistic discourse might be influenced by sleep. However, instead of using recognition, Experiment 2 indexed discourse memory via recall, which has been proposed to have a greater sensitivity to sleep-related memory effects (Berres & Erdfelder, 2021; Diekelmann et al., 2009; Lipinska et al., 2019). However, without the kind of recognition probes corresponding to the three levels of discourse representation, it is difficult, if not impossible, to separate out their individual quality. Therefore, in using recall-based procedures in Experiment 2, we stepped away from a tripartite view of discourse memory and took a less theoretically bound approach.

Here, we made use of both free and cued recall to test whether sleep influences the *quantity* and *quality* of discourse memory respectively. The free recall test was a near-replication of Aly and Moscovitch (2010), who showed that young adults ($N = 10$) recalled a greater number of story propositions after sleep (vs. wakefulness) (see Lau et al., 2018 for evidence from adolescents; although see Cohn-Sheehy et al., 2022). In this task, participants listened to two short stories from the Wechsler Memory Scale-III and subsequently recalled them in a free recall procedure.

The cued recall test was designed to capture the *quality* of discourse memory with a more nuanced scoring approach. There are two reasons why we believe this is important. We know from prior studies (e.g., Fisher & Radvansky, 2018; Sachs, 1974) as well as Experiment 1 that surface details of a text are quickly, if not completely, forgotten shortly after exposure. We also know that over time, memories for discourse tend to become increasingly gist-based and distorted (e.g., Bartlett, 1932; Brainerd & Reyna, 2004; Reyna et al., 2016). All these suggest that discourse memories are bound to undergo qualitative changes over time. In light of these, we developed a cued recall paradigm to test the effect of sleep on the quality of discourse memory, where participants were given sentence fragments and recalled the word they think appeared in the story. This procedure constrains the range of responses that participants are producing, allowing us to more precisely ascertain the nature of any errors by directly comparing a participants' response and the verbatim word in the story. As this is a novel paradigm, it is necessary to explain it in some detail. Consider this excerpt from one of the Fisher and Radvansky's (2018) stories:

> *Nevertheless, the detailed confessions that have been made public, including that of George Fields, make it difficult to believe as has been argued, that the whole story was invented by Steve Flett so that he could strengthen his position in the government of Pitman.*

After reading these stories, participants were given a cued recall task, in which they completed sentence fragments using the word they think appeared in the story e.g.,

> *…so that he could_____his position in…*

This task can be thought of as in some ways a mirror image of associate production in Gaskell et al. (2019), which motivated the episodic context account and the current set of experiments: Associate production taps memory for sentential context given the cue of a word (e.g., homonym) while this fill-in-the-blank task taps memory for a word given the cue of the sentential context. We used the responses generated by the participants to infer the quality of discourse representation in two separate analyses. We first directly examined surface memories, where we calculated the number of trials on which participants were able to reproduce the verbatim word used in the story (e.g., *strengthen*). This provided an all-or-nothing measure of memory. Second, to index more graded, qualitative changes to discourse memory, we classified each response based on the extent to which it fits with the story's event model, regardless of whether it was the verbatim word. This categorisation approach was developed with the help of 25 pilot participants, who completed the fill-in-the-blank task after reading the stories. With these responses, we explored the potential categories that could be derived. The first author and two research assistants individually examined the responses and independently proposed categories based on recurring patterns and themes observed in the data. Subsequently, through rigorous discussions and iterative refinement, we reached a

consensus on the final set of four categories (see Table 6).

We note here that while this graded (vs. all-or-nothing) approach allows us to capture the nature of any errors, the process of response categorisation is subjective in nature—a limitation in the existing literature that was highlighted in the introduction. As a remedy, we conducted an exploratory analysis, where we investigated the utility of a more objective scoring approach, namely Latent Semantic Analysis (Landauer & Dumais, 1997). To foreshadow our results somewhat, finding from this exploratory analysis aligned well with the subjective categorisation approach.

To sum up, Experiment 2 stepped away from a tripartite view of discourse memory and took a less theoretically bound approach in accessing discourse memory. Specifically, we used (i) free recall to test if sleep influences the retention of discourse propositions (i.e., quantity) and (ii) cued recall, referred to as fill-in-the-blank, to test if sleep influences how discourse memory changes qualitatively over time (i.e., quality).

Methods

Design overview

Experiment 2 comprises two sessions, separated by 12 h (see Fig. 6). Here, we made use of two sets of stories (vs. one in Experiment 1), each corresponding to a different outcome measure: Two short stories from the Wechsler Memory Scale-III were tested via free recall, while the four Fisher and Radvansky's (2018) stories used in Experiment 1 were tested via cued recall (i.e., fill-in-the-blank). In administering the former, we followed Aly and Moscovitch (2010) closely by using the same stories and presentation modality (i.e., auditory) but had three key changes: First, sleep vs. wake was manipulated within-participants in their study but between-participants in ours (in keeping with Experiment 1). Second, in their study, participants recalled each short story immediately after exposure, as well as 12 h later. While this enabled the researchers to take baseline difference into account, repeated testing without feedback (i.e., retrieval practice) has been shown to reduce the effect of sleep in other paradigms (e.g., Abel et al., 2019; Antony & Paller, 2018; Mak & Gaskell, 2023). We, therefore, tested one of the two stories at baseline in Session 1 and then both in Session 2, enabling us to simultaneously test for a "purer" effect of sleep and whether repeated testing is a modulating factor (Bäuml et al., 2014). Finally, participants in Aly and Moscovitch (2010) recalled the stories on the phone. As our experiment was conducted online over the COVID-19 pandemic, we had concerns over participants lacking recording equipment or producing low-quality recording; we, therefore, had participants recall the Wechsler short stories by typing them out.

For the Fisher and Radvansky stories, we followed Experiment 1 by using self-paced reading. However, contrary to Experiment 1, half of the stories were tested via fill-in-the-blank in Session 1, serving as baseline measures, while the other half were tested in Session 2. This was motivated by the finding that taking baseline measures into account can improve sensitivity to sleep-related memory effects (Berres & Erdfelder, 2021; Lipinska et al., 2019).

Participants were randomly assigned to the Wake or Sleep groups: In the Wake group, participants began Session 1 at 9AM (±1 hr) and Session 2 at 9PM (±1 hr) on the same day. Those in the Sleep group began Session 1 at 9PM (±1 hr) and the second session 12 h later (including a period of overnight sleep) at 9AM (±1 hr) the next day. Note that in contrast to Experiment 1, Experiment 2 dropped the Immediate control groups (i.e., Immediate-AM and Immediate-PM). This was because the current design was able to take participants' baseline performance in Session 1 into account, which helps to rule out time-of-day effects. In short, the study had one between-participant variable: Group (Wake vs. Sleep), with baseline performance in Session 1 serving as a within-participant covariate. All aspects of Experiment 2 were pre-registered ahead of data collection (https://aspredicted.org/P64_CQZ). Any deviations are explicitly noted.

Participants

Recruitment procedure was identical to Experiment 1. A total of 290 respondents from Prolific filled out a screening survey; 51 of them did not meet our inclusion criteria, leaving us with 239 respondents. Of these, 105 (70 females; M_{age} = 21.6; SD_{age} = 2.34) completed both sessions. Nine of them were excluded from further analysis: giving an SSS rating of 6 or more (N = 3), reporting to have taken part in Experiment 1B (N = 2), to have a nap before Session 2 in the Wake group (N = 1), and to have less than 6 h of sleep or poor sleep quality the night before (N = 3). The final sample size was therefore 96 participants, with 47 in the Wake group and 49 in the Sleep group. This sample size was informed by Fisher and Radvansky (2018), who also had 48 participants per group.

Materials

Short stories. Two short stories from the Logical Memory section of the Wechsler Memory Scale III (WMS-III; British version) were used. Stories A and B had 66 and 85 words respectively. Following Aly and Moscovitch (2010), the stories were presented in auditory form. The respective recordings, read aloud by a female native British English speaker with a Southern English accent, were 21 and 29 s in duration.

Long stories. The four stories from Experiment 1 were shortened and slightly modified to reduce reading time (M_{Exp1} = 621 words vs. M_{Exp2} = 500 words; see Appendix C3 for an example). Care was taken to ensure that the gist of the stories remained the same.

Selection of target words for the fill-in-the-blank task. Sixteen target words were chosen from each long story, hence 64 targets in total (see Appendix D for full list). According to the English Lexicon Project (Balota et al., 2007), these words have low-to-medium log frequency (M = 8.5 per million, SD = 1.6) and are low in concreteness (M = 2.51, SD = 0.66). Importantly, these words were chosen for their low contextual predictability, indexed by a norming study. Following the procedures in Nation and Snowling (1998), 20 native English speakers recruited via Prolific (who did not take part in the pilot or the main study) completed a cloze task without reading the stories. They were given phrases or sentences that had the target words removed (e.g., "…so that he could_____his position in…"). Each phrase/sentence contained exactly seven words, taken verbatim from the long stories.[4] The task instruction was to fill in the blank using the first word (single or hyphenated) that came to mind. Each trial showed one blank, and trial

Table 6
The four pre-registered categories[1] to which a response in the fill-in-the-blank task was assigned.

Categories	Descriptions	Examples
Verbatim/ Synonym	The verbatim word or close synonyms—words that map onto roughly the same proposition	*strengthen, improve*
Near Gist	Response had some semantic overlap with the verbatim word but does not contradict the story's event model	*maintain, keep*
Far Gist	Response share little semantic overlap with the verbatim word from but does not necessarily contradict the story's event model	*face, fill, use*
Contradiction	Response contradicts the story's event model	*reclaim, retake, fight*

[1] At the request of a reviewer, the first category was renamed to *Verbatim/Synonym* from *Alignment*, the second category to *Near Gist* from *Minor Distortion*, and the third category to *Far Gist* from *Major Distortion*.

[4] We chose 7 words because in another norming study with 5 participants, we found that increasing the length of the phrases to 10 substantially increased the predictability of the cloze. 7 appeared to be the 'right' length in that it does not give too much away but at the same time not too obscure.

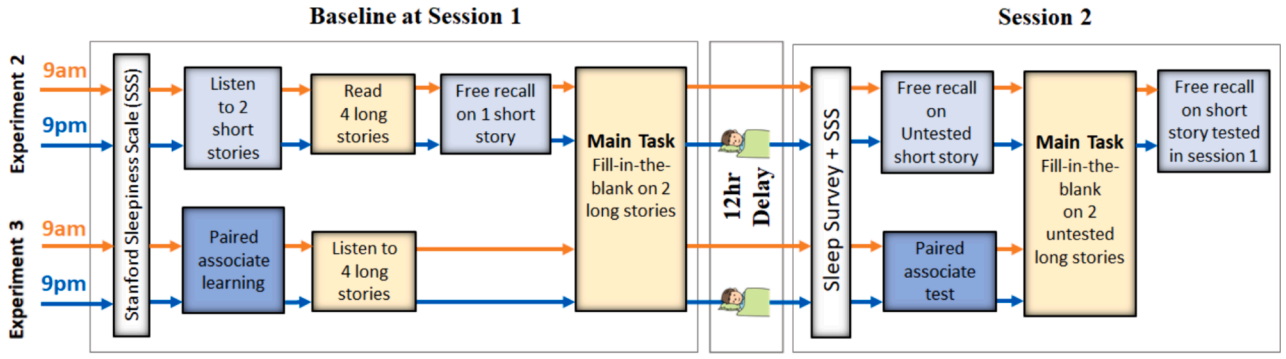


Fig. 6. Procedure in Experiments 2 and 3.

presentation was blocked by story, with the title of the story being displayed throughout each block. Block order was randomised, and so was the trial order within block. This norming study showed that the 64 target words could not be easily predicted by their surrounding contexts, as each was guessed by no more than 2 (of the 20) participants ($M = 1.2$, $SD = 0.4$).

Procedure

Session 1 began with participants giving a sleepiness (SSS) rating, followed by an audio check. Participants then listened to the two short stories from the WMS-III Memory Scale, each presented once in a random order. The instruction, taken verbatim from the WMS-III manual, was to listen to the stories carefully and to remember them the way they were presented. Afterwards, participants read the four long stories from Experiment 1, whose order of presentation was randomised. The reading instructions and procedure were identical to those in Experiment 1. Following this, participants completed a baseline free recall task on one of the short stories, where they were given a story title (e.g., Story A) and had 3 min to recall it by typing it out. Participants were instructed to use the "exact wordings" whenever possible. The next task was fill-in-the-blank at baseline, where participants were presented with a total of 32 sentence fragments from two randomly selected long stories. The task instruction was to fill in the blank using the word they thought appeared in the long stories. If unsure, an educated guess was welcomed.

Session 2 took place approximately 12 h after Session 1. It began with a survey, where participants gave an SSS rating and answered a sleep survey. Afterwards, participants had 3 min to complete free recall on the untested short story (e.g., Story B). This is referred to as the "Session 2-New" free recall. Following this was the fill-in-the-blank task based on the remaining two untested long stories. Task procedure was identical to that at baseline in Session 1. Finally, Session 2 ended with a free recall on the short story tested at baseline (e.g., Story A). This is referred to as the "Session 2-Repeated" recall, which allowed us to test whether a sleep effect may be affected repeated testing (e.g., Antony et al., 2017; Bäuml et al., 2014).

Results

Group characteristics

As summarised in Table 7, the Wake and Sleep groups were highly comparable in terms of various key characteristics (e.g., level of sleepiness, morningness/eveningness). Independent *t*-tests comparing SSS, morningness/eveningness scores, and number of intervening hours revealed no significant between-group differences (*ps* > .49).

Scoring

Short story free recall. Following the WMS-III scoring protocol, one point was awarded to each correctly recalled story proposition. For instance, if the story contains "He is a policeman", one point was awarded if participants recalled "He is a policeman" or "He is a cop". The maximum point for each story was 24. Two trained research assistants, blind to a participant's group allocation, completed the scoring individually. Inter-rater agreement rate was high at 98 %, and disagreements were resolved by a third rater.

Table 7
Group characteristics of the Wake and Sleep groups in Experiments 2 and 3. Values in parentheses indicate standard deviations.

Experiment	Experiment 2		Experiment 3	
Characteristics	Wake Group	Sleep Group	Wake Group	Sleep Group
N of participants	47	49	48	48
N of females	30	34	31	32
Mean SSS in Session 1 (Max = 7)	2.38 (1.45)	2.54 (1.63)	2.44 (1.51)	2.79 (1.98)
Mean SSS in Session 2 (Max = 7)	2.69 (1.88)	2.50 (1.44)	2.54 (1.22)	2.62 (1.58)
Mean Morningness/ Eveningness (Max = 25)	12.11 (3.45)	11.88 (3.88)	13.01 (3.83)	12.68 (3.28)
Mean *N* of hrs between study and test	11 hr 24 min (49 min)	11 hr 45 min (1 hr 2 min)	11 hr 36 min (58 min)	10 hr 52 min (1 hr 29 min)
Mean hrs of sleep	NA	7 hr 56 min (1 hr 38 min)	NA	7 hr 46 min (1 hr 26 min)

Notes. (1) SSS stands for Stanford Sleepiness Scale; it ranges from 1 to 6, with higher values indicating greater sleepiness. (2) Morningness/Eveningness score ranges from 5 to 25, with higher values indicating greater morningness preference.

Long story fill-in-the-blank. Participants' responses were corrected for any obvious spelling mistakes (defined as Levenshtein distance < 2). As outlined in the introduction, there were two dependent variables: [1] the number of verbatim words correctly recalled, and [2] the number of responses falling into each of the four pre-registered categories: Verbatim/Synonym, Near Gist, Far Gist, and Contradiction. The same research assistants, who were blind to a participant's group allocation, independently assigned each response to one of the four categories. Inter-rater agreement rate was satisfactory, at 80.2 % (agreement rate in each category: >75 %). Disagreements were resolved by discussion with a third rater.

Analysis approach

The analysis approach here differed from that in Experiment 1 due to differences in experimental design. Experiment 1 had the Immediate control groups (Immediate-AM & Immediate-PM) to account for time-of-day effects. In contrast, participants in Experiment 2 completed baseline measures for both free recall and fill-in-the-blank in Session 1, allowing us to assess time-of-day effects on encoding. A further advantage of this

is that we can more directly test the effect of sleep vs. wake by including baseline performance as a within-subject covariate. Statistically, this is more powerful than testing for an interaction between Group and Session.

Following our pre-registered plan, we used independent *t*-tests (or Mann-Whitney U tests if the assumption of normality is violated) to check whether time-of-day affected baseline performance in Session 1. Then, to compare performance between the Wake and Sleep groups in Session 2, we used analysis of covariance (ANCOVA), where the dependent variable was a participant's performance in Session 2 while the covariate was baseline performance in Session 1.

Confirmatory analyses

Short story free recall. At baseline, the Wake and Sleep groups recalled a comparable number of story units [M_{Wake} = 35.37 % (*SD* = 20.74 %) vs. M_{Sleep} = 37.41 % (*SD* = 17.95 %); *t*(90.89) = -0.51, *p* = .601], suggesting no time-of-day effects (see Fig. 7).

Next, in testing the effect of sleep vs. wake, two separate ANCOVAs were performed. The first had the number of story units recalled in Session 2-New as the dependent variable. It revealed no main effect of group, [M_{Wake} = 24.73 % (*SD* = 19.55 %) vs. M_{Sleep} = 23.97 % (*SD* = 15.83 %); *F*(1, 93) = 0.177, *p* = .674, η^2 = .002]. As for the second ANCOVA, the dependent variable was the number of recalled story units in Session 2-Repeated. Again, there was no main effect of group, [M_{Wake} = 31.83 % (*SD* = 19.9 %) vs. M_{Sleep} = 33.76 % (*SD* = 17.01 %); *F*(1, 93) = 0.03, *p* = .864, η^2 < .001]. Together, these suggest that sleep did not benefit the number of story propositions being recalled, regardless of repeated testing (without feedback). In other words, despite using the same stories and scoring protocol as Aly and Moscovitch (2010), we did not replicate their findings (or Bäuml et al., 2014).

Long story fill-in-the-blank. Focussing first on the verbatim responses (see left panel of Fig. 8): At baseline, participants in the Sleep group (who read the stories in the evening) recalled significantly fewer verbatim words (*M* = 13.1 %, *SD* = 9.5 %) than those in the Wake group (who read the stories in the morning) (*M* = 20.68 %, *SD* = 16.9 %), as indicated by a Mann-Whitney *U* test, *W* = 1462, *p* = .022, *r* = 0.234. This suggests that participants who read the stories in the evening (vs. morning) retained fewer surface details of the stories when tested soon afterwards. This finding, while not anticipated, is in line with the significant effect of Start Time in the textbase level in Experiment 1. Together, they suggest that reading strategies varied between morning and evening, with the latter tending to be less literal (Lorenzetti & Natale, 1996; Oakhill, 1986). We do not interpret this further as we did not have any *a priori* hypothesis regarding circadian differences.

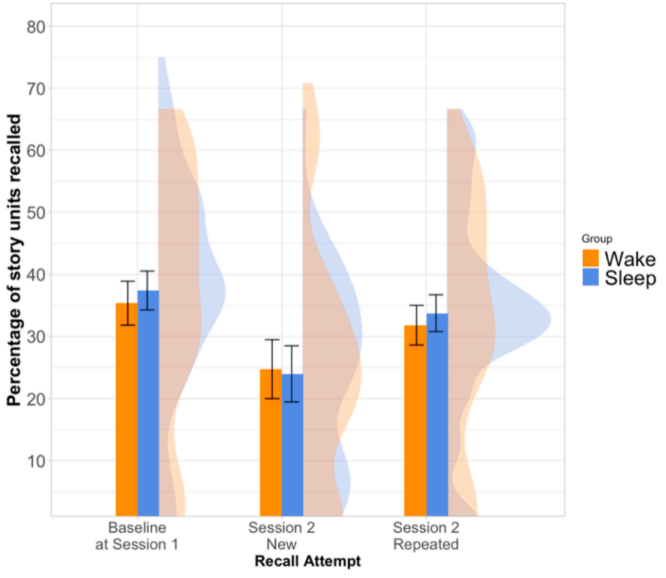


Fig. 7. Percentage of story units recalled in WMS-III story recall (Experiment 2), summarised across groups and recall attempts. Note. Error bars represent 95 % between-subject CIs while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

Next, regarding the effect of sleep on verbatim responses, an ANCOVA revealed no effect of group [M_{Wake} = 11.37 % (*SD* = 16.7 %) vs. M_{Sleep} = 10.91 % (*SD* = 11.8 %); *F*(1, 93) = 2.32, *p* = .131, η^2 = .017], indicating that the Wake and Sleep groups retained a similar number of verbatim words after a 12-hr delay.

Then, we turn to the subjective scoring approach where each response was assigned to one of the four predetermined categories: Verbatim/Synonym, Near Gist, Far Gist, Contradiction. Distribution to each category is summarised across groups and sessions in Fig. 9. Note that each participant produced a total of 32 responses in each session, so in Fig. 9, a mean of 50 % in a category means that participants on average produced 16 responses in that category.

M_{Sleep} = 38.45 % (SD = 13.6), M_{Wake} = 45.28 % (SD = 16.9); *t* (88.36) = -2.17, *p* = .032, *d* = -0.45]. This was driven by the sleep participants having poorer recall of the verbatim words as shown in the previous analysis. Apart from this comparison, the two groups did not differ significantly in their number of responses in the remaining categories, *t*s < 1.8, *p*s > .076, *d*s < 0.37.

Following our pre-registered analysis plan, we ran four separate ANCOVAs to evaluate the effect of sleep vs. wake on each of the four categories, with performance at baseline serving as covariates (see left panel of Table 8). We recognise a significant limitation inherent in this categorical/analysis approach, namely, that the distribution of responses across the four categories is not independent of each other, as each response was assigned to one of the four categories. Therefore, any potential sleep-wake difference within a category has the potential to be accompanied by a corresponding difference in the opposite direction within another category.

The Sleep group produced more Verbatim/Synonyms than the Wake group in Session 2 [M_{Sleep} = 34.38 % (SD = 16.15 %) vs. M_{Wake} = 32.31 % (SD = 19.32 %)], but this was not statistically significant [*F*(1, 93) = 2.61, *p* = .110, η^2 = .023]. There was a main effect of group in the Near Gist category [*F*(1, 93) = 4.58, *p* = .035, η^2 = .047] such that the Sleep (vs. Wake) group produced significantly more Near Gist in Session 2 [M_{Sleep} = 20.79 % (SD = 7.29 %) vs. M_{Wake} = 17.75 % (SD = 6.45 %)]. Furthermore, the Sleep group (*M* = 28.44 %, SD = 9.97 %) produced significantly *fewer* Far Gist than the Wake group (*M* = 32.64 %, SD = 12.45 %) in Session 2 [*F*(1, 93) = 4.27, *p* = .042, η^2 = .043].

Finally, although not pre-registered, we compared performance between baseline and Session 2, collapsed across groups, to index how discourse memory transformed over 12 h. Paired *t*-tests revealed that participants produced fewer Verbatim/Synonyms and more Near and Far Gist in Session 2 (vs. baseline) (*t*s > 2.35, *p*s < .021, *d* = 0.33–0.5). In contrast, the number of Contradictory responses was similar across sessions, *t*(95) = 1.37, *p* = .174, *d* = 0.18. Together, our findings suggest that after a 12-hr delay, discourse memory tended to become more distorted (although not in a contradictory way), but the extent of distortion was lower post-sleep (vs. post-wake). Specifically, compared to the wake group, more responses were categorised as Near Gist but fewer as Far Gist, consistent with the gist of the missing words being preserved after sleep.

Exploratory analyses

LSA The fill-in-the-blank task was scored by two independent raters who assigned each response to one of the four pre-registered categories. Although the inter-rater agreement rate was high at 80.2 %, it was far from perfect, and its subjective nature made it relatively difficult for future studies to replicate. Furthermore, since a response could only be assigned to one of the four categories, the distribution of responses is not

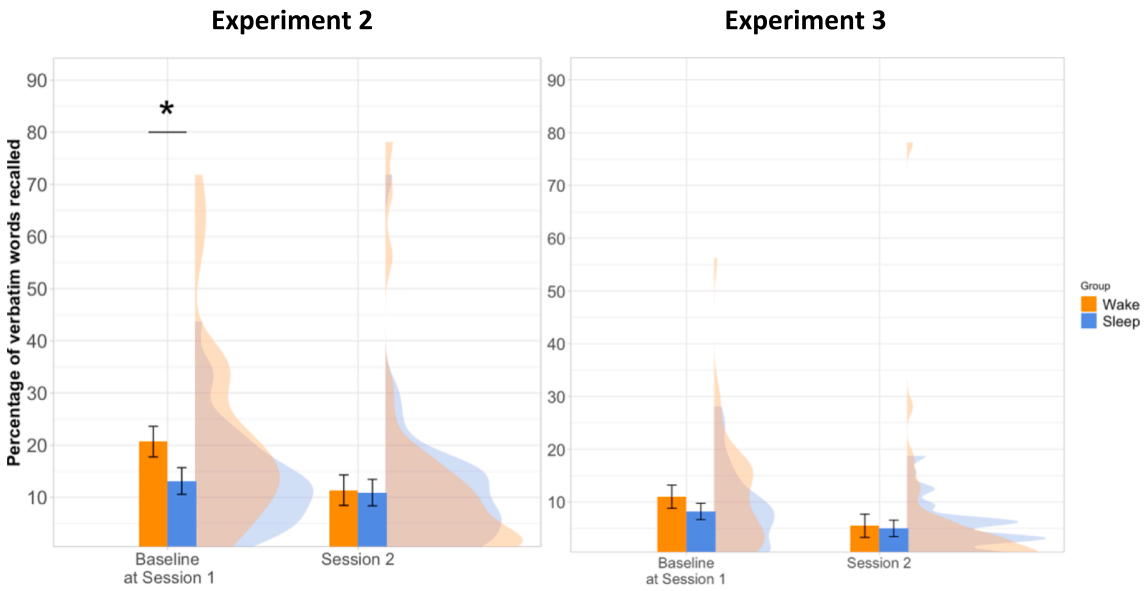


Fig. 8. Percentage of verbatim words recalled across groups and sessions in the fill-in-the-blank task in Experiments 2 (left) and 3 (right). Note. [1] * denotes statistical significance ($p < .05$). [2] Error bars represent 95 % between-subject CI while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

independent; this means that if the sleep (vs. wake) group had significantly more responses in one category, it must have fewer responses in the other(s). In light of all these limitations, we ran an exploratory analysis where we used Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer et al., 1998) to index the semantic distance between a participant's response and the target verbatim word, thereby giving a more objective measure of the degree of distortion.

Latent Semantic Analysis (LSA) was built upon the theoretical notion that words occurring in similar linguistic contexts have similar meaning (Sahlgren, 2008; Harris, 1954). It conceptualises word meanings as vectors in a high-dimensional semantic space, derived from word distribution in language corpora. With these vectors, one can calculate the likelihood with which two words occur in similar documents, embodied in an LSA-cosine value. Two words having a greater LSA-cosine value means they occur in more similar documents, and hence are referred to as being more semantically related (e.g., *strengthen-improve*: 0.44 vs. *strengthen-fill*: 0.09).

We reasoned that if participants in the Sleep (vs. Wake) group produced responses that were less distorted, their non-verbatim responses should share a higher LSA-cosine with the target words. To test this possibility, we calculated the LSA-cosines for target-response pairs, following the advice from the LSA Handbook (Dennis, 2014). Since we had 96 participants and 64 responses from each participant, we had a total of 6144 target-response pairs. We excluded i) 864 verbatim pairs (e.g., *strengthen-strengthen*) because we are interested in the degree of distortion, and ii) 238 pairs because the response words were non-existent in the LSA corpus. The exploratory analysis was based on the remaining 5042 target-response pairs.

To validate this LSA measure, we first compared it to the categorisation metric. The figure in Appendix E summarises the distribution of LSA-cosines across the four categories. Reassuringly, the mean LSA-cosine was highest in the Verbatim/Synonym category, followed by Near, Far Gist, and Contradiction. A one-way ANOVA confirmed a significant difference between these categories, $F(3, 2165) = 66.9$, $p < .001$, giving us confidence that the LSA metric is comparable to, albeit different from, the categorisation metric. We will revisit this point in General Discussion.

We then evaluated the effect of wakefulness vs. sleep on this LSA measure (see left panel of Fig. 10). At baseline in Session 1, there was no difference between the Wake and Sleep groups, $t(93.2) = 0.62$, $p = .535$. An ANCOVA with a participant's mean LSA-cosine in Session 2 as the dependent variable and mean LSA-cosine at baseline as the covariate revealed a main effect of group [$F(1, 93) = 6.86$, $p = .010$, $\eta^2 = .066$], such that the Sleep group produced responses that were closer to the target words in LSA semantic space[($M_{Wake} = 0.188$ (SD = 0.029) vs. $M_{Sleep} = 0.204$ (SD = 0.03)]. We interpret this finding as indicating that in Session 2, the non-verbatim responses from the Sleep group were less distorted than those from the Wake group, consistent with the findings from the categorisation approach.

Discussion

Motivated by the null findings from sentence recognition in Experiment 1, Experiment 2 indexed discourse memory using free and cued recall, which might be more suitable for detecting sleep-related memory effects (e.g., Diekelman et al., 2009). In free story recall, we did not replicate Aly and Moscovitch's (2010) finding that sleep (vs. wake) enhanced the number of story propositions being recalled, regardless of whether an immediate retrieval practice (i.e., baseline performance) was afforded. In contrast, our fill-in-the-blank task showed that sleep may have an effect on the *quality* of discourse memory such that the degree of time-related distortion was lower after sleep than after wakefulness. We will consider these findings in turn, beginning with free recall.

The reported effect size for sleep in Aly and Moscovitch (2010) was large, at Cohen's $d = 1.48$. This means that our sample size (i.e., 96 participants) gave us over 95 % statistical power (assuming alpha = 0.05) to detect this sleep benefit. However, we did not replicate the sleep benefit reported by Aly and Moscovitch's (2010) (or Bäuml et al.'s (2019), who showed a large sleep benefit in story recall when no retrieval practice was afforded). Our null finding mirrors that from recent studies (Cohn-Sheehy et al., 2022; Experiment 2; van Rijn et al., 2017), which found no sleep-dependent effect in free story recall among >90 young adults. Together, these null findings lead us to suspect that the effect size reported by Aly and Moscovitch (2010) might have been inflated as a consequence of their small sample size ($N = 10$). However, despite our experiment being a near-replication of Aly and Moscovitch (2010), differences in experimental design might have contributed to

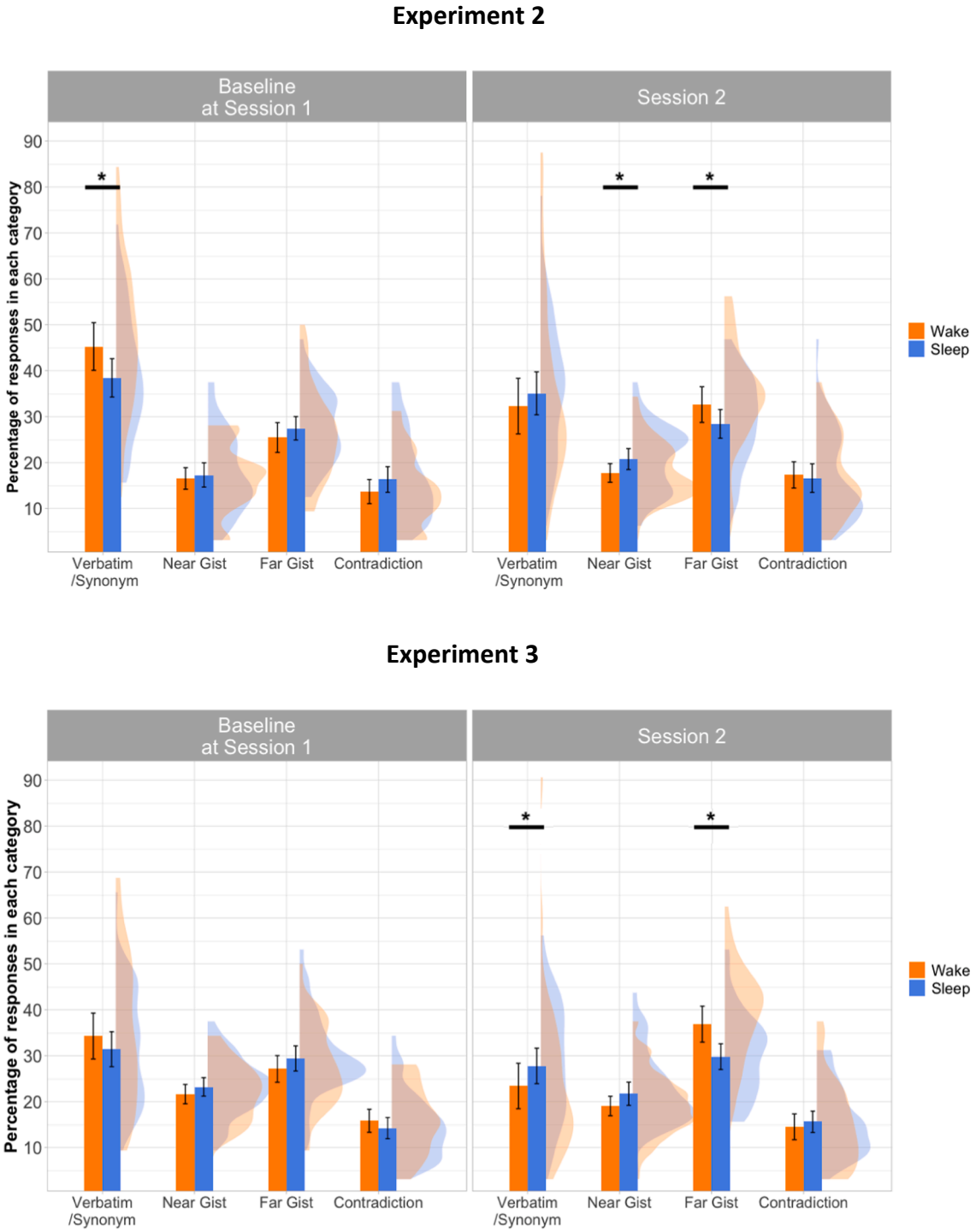


Fig. 9. Percentage of responses in each of the four categories across groups and sessions in the fill-in-the-blank task of Experiment 2 (top) and Experiment 3 (bottom). Note. [1] * denotes statistical significance ($p < .05$). [2] In comparing the two groups at baseline, we used independent t-tests, and in comparing the two groups in Session 2, we used ANCOVA, with performance at baseline as a covariate. [3] Error bars represent 95 % between-subject CI while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

the discrepant results. First, unlike Aly and Moscovitch (2010), our participants read four additional stories after listening to the short stories from WMS-III, which might have led to memory contamination. Another contributory factor might be the modality of the recall task: Participants in Aly and Moscovitch (2010) recalled the WMS-III stories verbally on the phone, while ours typed out the stories. Written recall might be more prone to a near-floor effect as typing is generally more effortful than speaking. Alternatively, the presence of the researcher on the phone may have induced more effort from the participants. In sum, the null findings from free story recall provide evidence against a strong version of the episodic context account and underscores the necessity for this framework to reassess the extent to which sleep influences discourse memory.

Moving on to the fill-in-the-blank task, we assessed participants'

Table 8
ANCOVA table summarising the effects of group (Wake vs. Sleep) on the number of responses in each of the four categories (Verbatim/Synonym, Near Gist, Far Gist, and Contradiction) in Experiment 2 (left) and 3 (right).

Categories	Experiment 2				Experiment 3			
	F	*p*	η^2	*Direction*	*F*	*p*	η^2	*Direction*
Verbatim /Synonym	2.61	.110	.023	–	5.63	.020*	.042	Sleep > Wake
Near Gist	4.58	.035*	.047	Sleep > Wake	2.1	.15	.022	–
Far Gist	4.27	.042*	.043	Sleep < Wake	11.21	.001*	.109	Sleep < Wake
Contra-diction	0.63	.431	.010	–	0.38	.537	.004	–

Note. * denotes statistical significance ($p < .05$).

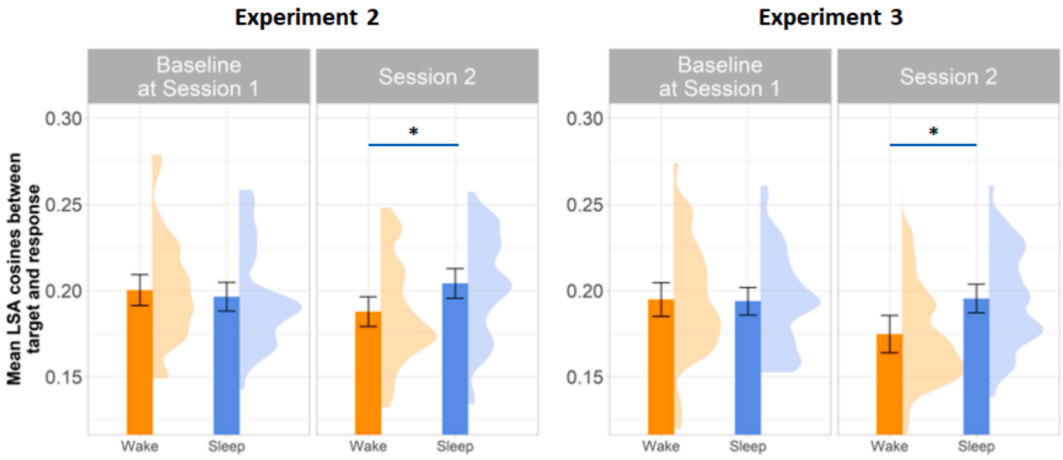


Fig. 10. Mean target-response LSA-cosines across groups and sessions in Experiment 2 (left) and Experiment 3 (right).Note. [1] * denotes statistical significance at p < .05. [2] Error bars represent 95 % between-subject CI while the density functions represent the distribution of the data. For the density plot, we used the stat_halfeye function in the ggdist package, which passes the adjust parameter (set to 0.8) to the density function that uses a Gaussian kernel by default.

performance using two approaches. The first is whether the participants were able to recall the verbatim words used in the story. Unsurprisingly, participants in both groups were performing at near-floor in Session 2. This mirrors the sentence recognition data in Experiment 1, where the A′ score for the surface level was hovering slightly above chance in Session 2. Together, these suggest that discourse memory on the surface level is perhaps too impoverished for sleep to act on.

Then, for the second scoring approach, we classified each participant's response into one of four pre-registered categories: Verbatim/Synonym, Near Gist, Far Gist, and Contradiction. This enabled us to examine qualitative changes to discourse memories, which are known to become increasingly distorted over time (e.g., Bartlett, 1932). We found that while there were more distortions in Session 2 (vs. baseline in Session 1), participants in the Sleep (vs. Wake) group produced more Near Gist but fewer Far Gist responses. These findings suggest while the distribution of responses shifted towards greater distortion 12 h later, the degree of distortion was lower if sleep soon followed story exposure. To put this another way, the Sleep group seems to have shown better gist preservation than the Wake group. Similarly, our exploratory analysis using LSA showed that in Session 2, non-verbatim responses from the Sleep (vs. Wake) group had higher LSA-cosines with the target words that appeared in the story, suggesting a lower degree of distortion after sleep. There are at least two plausible explanations for these findings. First, sleep might have actively consolidated the gist of the stories, thereby reducing the degree of distortion. Alternatively, sleep might have passively protected discourse representations against interference. We will return to these points in greater detail in General Discussion. Finally, we end this section by noting that although our categorisation and LSA measures provided converging results, the latter was exploratory in nature. It is therefore necessary to interpret the findings accordingly (e.g., Bishop, 2020). In light of this, we decided to perform a confirmatory replication of Experiment 2.

Experiment 3

The fill-in-the-blank data from Experiment 2 suggest that sleep may influence the quality (but not necessarily the quantity) of memory derived from naturalistic discourse such that the degree of time-related distortion may be lower after sleep than after wakefulness. This finding was supported by both the subjective categorisation and the more objective LSA approach, the latter of which was exploratory in nature. Here, we set out to replicate these findings in a confirmatory protocol, with two modifications: (1) Since the free story recall in Experiment 2 showed no sleep benefit, it was dropped and replaced by the paired-associate learning task from Experiment 1B. This was intended to provide us with a positive control and to keep Experiments 2 and 3 comparable in terms of duration and cognitive demand. (2) The long stories from Experiment 2 were presented aurally here, as opposed to visually. This was intended to test if the findings from Experiment 2 generalise to when discourse memory is acquired via the auditory domain. All aspects of Experiment 3 were pre-registered ahead of data collection (https://aspredicted.org/P3W_JBS). Any deviations are explicitly noted.

Methods

Experiment 3 employed the same design and procedure as Experiment 2 (see the lower half of Fig. 6 for visualisation), except free story recall was replaced by the paired-associate learning task from Experiment 1B. For the long stories, they were recorded using a female voice generated by a life-like text-to-speech software (Google Speech Services). This ensured that no specific words stood out and that acoustic features were equivalent across stories. Each recording lasted

approximately 3.5 min and could not be paused. Participants listened to the four stories once each, presented in a random order. This means that each word was heard exactly once, unlike Experiment 2 where participants could re-read a word multiple times before moving onto the next clause. The associated fill-in-the-blank task remained visually presented and was identical to that in Experiment 2.

Participants

Recruitment procedure was the same as previous experiments. After screening out those ineligible and those who dropped out, a total of 104 participants (67 females, 37 males; M_{age} = 22.1, SD_{age} = 2.56) from Prolific completed both sessions. Eight of them were excluded for meeting our exclusion criteria: giving an SSS rating of six or above (N = 2), reporting to have a nap between sessions in the Wake group (N = 3), to have less than six hours of sleep or to have poor sleep quality in the Sleep group (N = 3). The final sample size was 96 participants, evenly split between the Wake and Sleep groups.

Results

Group characteristics

As summarised in the right hand side of Table 7, the two groups were well matched in terms of various key characteristics (e.g., level of sleepiness, morningness/eveningness). Independent t-tests comparing SSS and morningness/eveningness scores revealed no significant between-group differences ($ps > .64$).

Confirmatory analysis

Paired-associate learning. The mean percentage of correct recall across groups is summarised in Table 9.

At baseline in Session 1, the Wake and Sleep groups recalled a similar number of word pairs [$t(94.96) = 1.02$, $p = .310$]. A one-way ANCOVA, with the number of correct recalls at baseline as the within-subject covariate confirmed that the Sleep (vs. Wake) group recalled significantly more word pairs in Session 2 [$F(1, 95) = 10.44$, $p < .001$, $\eta^2 = 0.1$]. Therefore, replicating prior lab-based studies (e.g., Plihal & Born, 1997) and Experiment 1B, we observed a sleep-associated benefit, providing evidence that within this specific sample, it is possible to detect sleep-related memory effects, even when participants completed the study remotely and unsupervised.

Fill-in-the-blank. Three participants (1 Wake + 2 Sleep) were excluded from this analysis as they failed to submit a response to over 25 % of the trials due to a technical problem. This exclusion was unforeseen, and hence not pre-registered. However, whether or not these participants were excluded did not change the interpretation of the analyses below (N = 93).

Verbatim measure. The data are summarised in the right panel of Fig. 8. Baseline performance in Session 1 showed that the Wake (vs. Sleep) group recalled more verbatim words [M_{Wake} = 11 % (SD = 10.56 %) vs. M_{Sleep} = 8.2 % (SD = 7.37 %)}; however, contrary to Experiment 2, this was not statistically significant, as indicated by a Mann-Whitney U test (W = 1314.5, p = .229, r = 0.12). In Session 2, both the Wake and Sleep groups performed near floor, recalling on average 5.5 % (SD = 12.37 %) and 5.01 % (SD = 4.81 %) verbatim words respectively. An ANCOVA also revealed no effect of group [$F(1, 91) = 0.33$, $p = .567$, η^2 = .003].

Categorisation measure. Distribution to the four categories (Verbatim/Synonym, Near Gist, Far Gist, Contradiction) is summarised across groups and sessions in Fig. 9. Focusing first on performance at baseline in Session 1: The two groups did not differ significantly in their number of responses across all the four categories, $ts < 1.49$, $ps > .139$, $ds < 0.31$.

Turning to the effect of sleep: Four separate ANCOVAs, one on each of the four categories, were performed (see right panel of Table 8). In Session 2, Group had a main effect for Verbatim/Synonym [$F(1, 91)$ = 5.63, $p = .020$, $\eta^2 = .042$] such that the Sleep (M = 27.79 %, SD = 12.85 %) group produced significantly more Verbatim/Synonyms than the Wake group (M = 23.44 %, SD = 15.97 %). This comparison was not significant in Experiment 2, although it patterned in the same direction (i.e., Sleep > Wake). Similar to Experiment 2, the number of Near Gists was greater in the Sleep (M = 21.74 %, SD = 8.28 %) than in the Wake group (M = 19.09 %, SD = 7.01 %), although this was not statistically significant [$F(1, 91) = 2.10$, $p = .15$ $\eta^2 = .022$]. Critically, we replicated the main finding from Experiment 2 such that the Sleep (M = 29.81, SD = 9 %) group produced significantly fewer Far Gist responses than the Wake group (M = 36.89 %, SD = 12.27 %) in Session 2 [$F(1, 91)$ = 11.21, $p = .001$, $\eta^2 = .109$]. Finally, in line with Experiment 2, the two groups did not differ significantly in the number of Contradictory responses in Session 2 [$F(1, 91) = 0.38$, $p = .537$, $\eta^2 = .004$].

Finally, although not pre-registered, we compared performance between baseline and Session 2, collapsed across groups, to test whether discourse memory became more distorted over 12 h. In line with Experiment 2, paired t-tests revealed that participants generally produced fewer Verbatim/Synonyms [$M_{Baseline}$ = 32.84 % (SD = 14.37 %) vs. $M_{Session2}$ = 25.66 % (SD = 14.55 %), $t(93)$ = -4.96, $p < .001$, d = -0.50] but more Far Gist responses in Session 2 than at baseline [$M_{Baseline}$ = 28 % (SD = 9.5 %) vs. $M_{Session2}$ = 33.28 % (SD = 11.24 %), $t(93)$ = 3.58, $p < .001$, d = 0.5], indicative of discourse representations becoming more distorted over time. Contrary to Experiment 2 though, the number of Near Gists was significantly *lower* in Session 2 than at baseline [$M_{Baseline}$ = 22.47 % (SD = 6.42 %) vs. $M_{Session2}$ = 20.45 % (SD = 7.8 %); $t(93)$ = -2.20, p = .031, d = 0.28]. And consistent with Experiment 2, the number of Contradictions was similar across sessions [$M_{Baseline}$ = 14.06 % (SD = 7.96 %) vs. $M_{Session2}$ = 14.93 % (SD = 8.34 %); $t(93) = 0.807$, $p = .422$, $d = 0.11$].

To summarise, in line with the findings of Experiment 2, the distribution of responses shifted away from verbatim/synonyms and towards more distortions following a 12-hr delay; importantly, however, the degree of distortion was lower after sleep (vs. wake), as reflected by the Near and Far Gist categories. To put this another way, our findings suggest better gist preservation after sleep (vs. wake). However, the shift in distribution after 12 h is not identical between Experiments 2 and 3. Specifically, while there were fewer Far Gist responses after sleep (vs. wake) in both Experiments, this was accompanied by more Near Gist responses in Experiment 2 but more Verbatim/Synonyms in Experiment 3, perhaps reflecting a less substantial shift towards distortion for the sleep group in Experiment 3 compared to Experiment 2. If anything then, this suggests that the benefit of sleep for gist preservation was stronger in Experiment 3 than in Experiment 2.

Latent Semantic Analysis. Following the LSA protocol in Experiment 2, we excluded 463 (or 7.8 %) verbatim and 330 (or 5.5 %) responses that were non-existent in the LSA corpus. The right panel of Fig. 10 shows a participant's mean LSA-cosine, summarised across groups and sessions. At baseline in Session 1, there was no significant difference between groups, $t(87.81) = 0.16$, $p = .875$. Then, a one-way ANCOVA controlling for baseline performance revealed a main effect of group [$F(1, 91)$ = 9.87, $p = .002$, $\eta^2 = .095$] such that participants in the Sleep group produced responses that were closer to the target words in LSA semantic space [M_{Wake} = 0.175 (SD = 0.036) vs. M_{Sleep} = 0.195 (SD = 0.028)]. This finding is in alignment with that from the categorical approach and confirms the exploratory finding from Experiment 2.

Table 9
Summary of cued recall performance in paired-associate learning (Experiment 3).

Group	Mean % of correct recall in Session 1 at baseline (SD)	Mean % of correct recall in Session 2 (SD)	% Changes across session
Wake	50.81 (21.1)	57.70 (20.7)	+6.89
Sleep	46.68 (19.0)	62.91 (18.2)	+16.23

Exploratory analyses

Following reviewers' requests, we re-ran all the analyses from Experiments 2 and 3 but dropped baseline performance in Session 1 as a covariate (see Appendix F for a summary). Nearly all significant sleep-wake comparisons from the pre-registered analyses turned non-significant; the only exceptions were the LSA analyses in both experiments and the Far Gist category in Experiment 3.

Discussion

Using the same fill-in-the-blank task as Experiment 2 but presenting the stories aurally, Experiment 3 set out to replicate the finding that time-related distortion to discourse memory would be of a lower extent after sleep (vs. wake). In line with Experiment 2, we found that participants in the Sleep (vs. Wake) group produced fewer Far Gist responses in Session 2, and their mean LSA scores in this session were also greater than the Wake group. The reduction in Far Gist responses seen in both experiments was balanced out predominantly by Verbatim/Synonyms in this Experiment and by Near Gist in Experiment 2. Together, these provided evidence that the degree of distortion was lower among those who had a sleep opportunity, once again suggesting that the effect of sleep on discourse memory may be more pronounced on a qualitative (vs. quantitative) level.

Although the results between Experiments 2 and 3 were highly comparable, they do differ in a few respects. Speculatively, this might be attributed to the stories being auditorily (vs. visually) presented. In Experiment 3, it was not possible to pause or replay the story recording, so if participants misheard a word or lost attention for a fleeting moment, memory representation for the stories could be compromised. Such representations might therefore be less robust compared to when the stories were read in a self-paced manner (Experiment 2). This has two implications: First, surface memory should be poorer in auditory (vs. visual) presentation, which was indeed the case—verbatim recall was worse in Experiment 3 than 2, regardless of session or group (e.g., $M_{\text{Exp2; Baseline}} = 16.81$ % vs. $M_{\text{Exp3; Baseline}} = 9.56$ %; $W = 2940$, $p < .001$). Second, since surface memory was compromised, this entails a knock-on effect on the higher levels of discourse memory, suggestive of more distortions in Experiment 3 (vs. 2). This appeared to be the case; for example, the mean number of Near Gist at baseline of Experiments 2 and 3 are 16.75 % and 22.47 % respectively, $t(179.91) = -5.37$, $p < .001$. Worthy of note here is that the effect size for the wake-sleep comparison increased markedly from Experiment 2 to 3; for example, this increased from $\eta^2 = 0.043$ to 0.109 in the Far Gist category and from $\eta^2 = .066$ to.095 in the LSA analysis. This implies that sleep may have exerted a more prominent effect when the initial discourse representation was encoded at a lower strength (but not at floor). This proposal fits with (1) the finding that there was a less substantial shift towards distortion for the sleep group in Experiment 3 compared to Experiment 2, and (2) prior evidence suggesting that sleep-associated benefits tend to be stronger in memories that are weakly (but not poorly) encoded (Drosopoulos et al., 2007; Denis, Mylonas, et al., 2021; Schapiro et al., 2017; cf. Petzka et al., 2021). We will revisit this point in General Discussion.

General discussion

The episodic context account (Gaskell et al., 2019) postulates that new episodic memories of sentences are routinely formed during language comprehension. These representations can then be exploited in subsequent linguistic interactions, alongside long-term linguistic knowledge. The theory was originally developed to account for word-meaning priming effects (Rodd et al., 2013), by which encountering a lexically ambiguous word in a particular context would influence its subsequent interpretation 20 min or more later. Gaskell et al. (2019) showed that these priming effects declined over wake but were sustained and stabilised over sleep, consistent with episodic memory consolidation. Subsequent studies have shown that such priming effects are not restricted to classical lexical ambiguity (Ball et al., 2024; Curtis et al., 2022), and that the pattern of preservation of priming across sleep but not wake is a more general one (Mak, Curtis, et al., 2023). What is less clear is whether these contextual memories also have a role to play in the development and maintenance of general memory for discourse, potentially by underpinning the construction or retention of event models (e.g., Altmann & Ekves, 2019; Graesser et al., 1997). The current research, therefore, tested a key prediction of the episodic context account that memory for naturalistic discourse should be enhanced after sleep (vs. wakefulness). From a sleep and memory perspective, this prediction may appear unsurprising as general accounts of sleep (e.g., Rasch & Born, 2007; Yonelinas et al., 2019) would make the same prediction; however, from a language comprehension perspective, this is a unique and novel prediction. As far as we are aware, all existing theories on language comprehension are mute on the effects of sleep. While some models (e.g., Blank et al., 2016) acknowledge the role of episodic memory networks, they do not explicitly address sleep. Our research bridges this gap by integrating sleep and discourse memory with comprehension, marrying the literatures on comprehension and sleep/memory. In other words, the episodic context account makes its primary theoretical contribution to language comprehension. As the prediction that sleep influences discourse memory is not specific to the episodic context account, it is not surprising that a few studies in the memory literature have tested whether sleep influences memory for discourse using free recall (and neutral stories); unfortunately, however, they yielded inconsistent results (e.g., Aly & Moscovitch, 2010; Cohn-Sheehy et al., 2022). In three experiments, we used both recognition- and recall-based paradigms to index the retention of discourse memory over a period of overnight sleep and daytime wakefulness. In doing so, we provided arguably the most comprehensive examination to date of how sleep may influence discourse memory—both quantitatively and qualitatively. Below, we briefly summarise our findings.

Experiment 1 followed the discourse processing literature by conceptualising discourse memory at three different levels: surface, textbase, and event model (Kintsch, 1988). To see how they might be differentially affected by sleep, we adopted a well-established sentence recognition paradigm (Schmalhofer & Glavanov, 1986). Consistent with Fisher and Radvansky (2018), we found that the three levels of representation were forgotten at different rates over 12 h, with surface and textbase memories, but not event model, showing a significant decline. However, contrary to the prediction of the episodic context account or other general account of sleep (e.g., Rasch & Born, 2007), we found no evidence that sleep was involved in the retention of any of the three levels, at least not when discourse memory was assessed via this recognition paradigm. In contrast, we found a clear sleep benefit in the positive control task of Experiment 1B (i.e., paired-associate learning). In the light of this dissociation and existing evidence suggesting that recall- (vs. recognition-) based procedures may have greater sensitivity to sleep-related memory effects (e.g., Berres & Erdfelder, 2021; Lipinska et al., 2019), we decided to assess discourse memory via recall in the subsequent experiments.

In Experiment 2, we first used free recall for two short stories from the Wechsler Memory Scale-III, which was a near-replication of Aly and Moscovitch (2010). Contrary to their finding, our participants recalled a similar number of story propositions regardless of whether they had a sleep or an immediate retrieval opportunity. This null finding suggests that the effect of sleep on schema-consistent discourse memory may be difficult to detect on a quantitative level. In addition to free story recall, participants in Experiment 2 also read longer stories and then completed a novel fill-in-the-blank task (a cued recall procedure), designed to capture the *quality* of discourse memory by comparing a participant's response word and the verbatim word in the story. Here, we found that while discourse memory showed decay/distortion over 12 h, the degree was lower if participants had a sleep opportunity. This finding converged with an exploratory Latent Semantic Analysis, which showed

that non-verbatim responses from the Sleep (vs. Wake) participants were more semantically related to the verbatim words used in the stories. These findings were then replicated in a confirmatory protocol in Experiment 3, where the stories were presented auditorily rather than visually. We interpreted these findings as indicating that sleep may affect discourse representation on a qualitative level, potentially by reducing distortion (or to put it another way, by better preserving the gist).

Overall, the findings from our series of experiments were more nuanced than predicted; they suggest that the effect of overnight sleep (vs. daytime wakefulness) on discourse memory is relatively modest and may depend on how it was assessed, potentially due to the involvement of other cognitive systems that do not rely on sleep to remain effective over time. Therefore, our null findings rule out a strong version of the episodic context account which prescribes a highly pervasive role to sleep in discourse memory. However, findings from cued recall (fill-in-the-blank) provide some support for a nuanced episodic context account such that sleep does play some role in the maintenance of discourse quality. Here, we flesh out our interpretation of the results, considering first the difference between recall and recognition, before turning to the effect of sleep.

Retrieval processes and sleep for discourse memory

Across our experiments, we made use of recognition, free, and cued recall to tap declarative memory for naturalistic, schema-consistent stories. Interestingly, sleep-related effects were only seen in cued recall (i.e., fill-in-the-blank in Exps 2 & 3). To determine why this was the case, we first consider the distinction between recall and recognition before turning to that between free and cued recall.

Recall vs. Recognition. Recent meta-analyses have shown greater sleep-related memory effects in recall than in recognition (Berres & Erdfelder, 2021; Lipinska et al., 2019), potentially because recall and recognition rely on distinct neural and cognitive mechanisms. On a neural level, recall is primarily hippocampus-dependent (e.g., Baddeley et al., 2001; Girardeau et al., 2017) while recognition is supported by extra-hippocampus regions, such as frontal-subcortical circuitry (e.g., Bastin et al., 2004; Bayley et al., 2008; Davachi et al., 2003; Mayes et al., 2002; Squire & Dede, 2015; Vargha-Khadem et al., 1997). Existing theories arguing for an active role of sleep in consolidation postulate that sleep is particularly relevant to hippocampus-dependent memories (see Paller et al., 2021 for a review), suggesting that sleep-associated memory effects may be more robust in recall than in recognition (Diekelmann et al., 2009; Wagner et al., 2007).

Free vs. Cued recall. Surprisingly, our study did not find evidence of sleep benefiting free recall of words or stories. However, we observed clear sleep-related effects in cued recall (i.e., fill-in-the-blank and paired-associate learning). Below we speculate why cued recall may be more sensitive than free recall in detecting sleep-related memory effects in the context of discourse memory.

Memory for naturalistic discourse may be considered associative in nature such that various discourse elements (e.g., characters, causality) are bound together to create a cohesive representation. Our cued recall task explicitly and specifically tested for such discourse bindings across every part of a story and forced participants to give a response to each blank. In contrast, during free recall, participants could give up retrieval anytime they liked, due to factors such as low effort/motivation. If, for example, a participant gives up retrieval after recalling the first sentence, it will leave discourse bindings in the middle and end of the stories completely untested, making free recall a relatively incomprehensive measure of discourse memory. Additionally, some discourse bindings may be too weak to be recalled during free recall but could be more successfully retrieved in cued recall due to the presence of prompts (e.g., sentence fragments); this may make free recall a less effective method for capturing the full extent and quality of discourse bindings. In other words, sleep-related memory effects in discourse memory may be less likely to emerge in free recall because it is less able to tap into discourse bindings as comprehensively and as consistently as cued recall.

Furthermore, sleep may primarily impact memories of schema-consistent stories on a qualitative, rather than a quantitative, level. Compared to arbitrarily paired words or schema-inconsistent stories, elements within schema-consistent stories usually have strong associative links due to the presence of e.g., causality (Radvansky, 2012). These strong links may leave little room for subsequent sleep to boost the quantity of such links; instead, it may help maintain their quality. The graded scoring protocols of fill-in-the-blank might have allowed us to better tap into such qualitative aspects than the all-or-nothing scoring protocol of free recall. In sum, the different nature of free and cued recall, along with their respective scoring protocols, may have contributed to the observed sleep-related effect in one task but not the other.

Finally, we note that the WMS-III stories used for free recall were always encoded before the long stories for fill-in-the-blank. This means that memories for the former might have been interfered by the latter, impacting the ability of sleep to benefit retention. Clearly, the sleep effect in paired-associate learning survived any potential interference from the long stories (Experiments 1B and 3), but if the sleep effect on stories is more subtle, then the ordering could be a contributing factor to why sleep had no detectable effect on free story recall.

Where does a sleep effect lie in the tripartite model?

Our Experiment 1 was built on the tripartite model of discourse processing, which views discourse memory as three interdependent representations—*surface*, *textbase*, and *event model* (e.g., van Dijk et al., 1983; Zwaan & Radvansky, 1998). Using the well-established sentence recognition paradigm, Experiment 1 found no evidence of sleep-related effects on any of the three levels, but relied on recognition memory as discussed above. The fill-in-the-blank task of Experiments 2 and 3 was not set up to readily tease apart the three levels of discourse representation as in sentence recognition. Regardless, it is useful to consider how the findings from fill-in-the-blank may be explained in relation to the tripartite model.

Verbatim recall/Surface Memory. Recall of verbatim words was consistently near-floor immediately after story exposure, with participants recalling on average 3 to 6 of the 32 verbatim words. This near-floor performance suggests that surface details were rapidly, although not entirely, forgotten soon after story exposure (Sachs, 1967; Fisher & Radvansky, 2018). This implies little room for post-encoding sleep to exert an influence. Therefore, any sleep-related effects on discourse memory are unlikely to be surface-level effects. Below, we consider findings from the subjective categorisation approach.

Verbatim/Synonym. Responses classified as Verbatim/Synonym are either the verbatim words or words that map roughly onto the same proposition. Given this, we believe it is reasonable to consider this as a proxy to the quality of textbase representation, such that the more the Verbatim/Synonyms, the better one's textbase memory. In Experiment 3, the Sleep group showed significantly more Verbatim/Synonyms, and though not statistically significant in Experiment 2, the pattern was similar. Together, they hint at the possibility of better textbase representation post-sleep (vs. post-wake). However, our data also suggest that whether a sleep-related effect is detected may depend on the initial quality of discourse memory: Relative to self-paced reading (Experiment 2), story listening (Experiment 3) resulted in poorer initial memory (as indexed by the numbers of verbatim recall and distortion responses). The fact that we found a sleep-related effect in Verbatim/Synonym when initial memory was of weaker (but not at-floor) quality coincides with evidence suggesting that such memory may be in more need of being consolidated over sleep, and hence more likely to be prioritised for sleep-related consolidation (e.g., Denis, Dipierto, et al., 2021; Drosopoulos et al., 2007; Payne et al., 2010).

Near and Far Gist. These categories do not have a one-to-one

correspondence with the three levels in the tripartite model. This is because not only are the three levels nested within each other, but they also have a reciprocal relationship. For example, a distorted event model may lead to a distorted textbase representation, and vice versa. Given this, it is difficult, if not impossible, to ascertain whether a Near/Far Gist response reflects a loss in textbase and/or event model memory. Therefore, the consistent finding that participants produced fewer Far Gist responses post-sleep warrants three interpretations: the quality of (i) textbase memories, (ii) event model, or (iii) both were less compromised in the Sleep (vs. Wake) group. Having said that, our data seem to favour interpretations (i) or (iii):

First, memory representations for event models tend to be the most resistant to forgetting, with evidence suggesting that they can remain robust even months after initial exposure (Doolen & Radvansky, 2021; Fisher & Radvansky, 2018). Assuming that sleep-related memory effects are constrained by the initial encoding strength (e.g., Denis, Mylonas, et al., 2021; Drosopoulos et al., 2007; Payne et al., 2010), this would predict that representations for event models—which are strong to begin with—are unlikely to benefit from a night of sleep. Second, in both Experiments 2 and 3, the number of Verbatim/Synonym was numerically or significantly higher post-sleep than post-wakefulness. If we take Verbatim/Synonym as a proxy to textbase memory, it suggests that the post-sleep reduction in Far Gist is at least partially driven by better textbase representation after sleep (vs. wake). However, as stated, the Near/Far Gist categories are unlikely to be pure measures of textbase memory or event model, as these discourse levels are interlinked. Therefore, our interpretation here is speculative in nature, and future research is needed to investigate how to tease apart the three levels of discourse representation in recall-based paradigms.

Contradiction. Contradictory responses violate a story's event model, and in both Experiments, there was no change in the number of Contradictions between baseline and Session 2. This coincided with prior findings that memories for event models tend to be stable over time (Fisher & Radvansky, 2018), suggesting that it might be possible to take Contradiction as a proxy to the quality of event model, such that the more Contradiction, the worse one's event model. Across the two experiments, there was no evidence that the number of Contradictions was affected by a night of sleep. Perhaps, this is because initial memory strength on this level was "too strong" for sleep to exert an effect.

To sum up, with reference to the tripartite model of discourse processing, data from our fill-in-the-blank task seem to suggest that a period of overnight sleep (vs. daytime wakefulness) may help maintain the quality of textbase memory (i.e., memories for the propositions irrespective of wording). However, this effect may be influenced by the initial strength of encoding such that a sleep benefit may be more detectable/pronounced when encoding strength is lower (but not at floor). As for the other two levels of representation, surface and event model, there is no evidence that sleep was involved in their maintenance, and following the encoding strength argument, surface memory is perhaps too impoverished for sleep to act on, and event model is perhaps too strong/stable for sleep to be a factor.

Returning to the episodic context account, it contends that episodic memory contributes to discourse processing by binding discourse elements together, resulting in an episodic representation that may be gist-like in nature (Curtis et al., 2022; Mak, Curtis, et al., 2023). Our fill-in-the-blank task provided some evidence for this and suggests that any sleep-related effects on discourse memory are likely to lie on the more abstract levels (i.e., textbase and/or event model). However, it is important to note that the results presented can also be explained within the broader framework of general theories on sleep-related memory effects (e.g., Rasch & Born, 2007). Further studies and theoretical advancements are necessary to better understand the specific conditions and mechanisms through which sleep impacts discourse memory.

Neurocognitive effects of sleep on discourse memory

Active consolidation accounts of sleep postulate that memory representations may be replayed within the hippocampus during sleep, leading to their stabilisation and strengthening (e.g., Born & Wilhelm, 2012). These accounts, therefore, predict that discourse memories should be enhanced post-sleep, either quantitatively and/or qualitatively. Our findings provide some support for this prediction by showing that sleep may help preserve the quality of discourse memory. However, it is also possible that offline consolidation during sleep, instead of strengthening discourse memories, may help tease them apart and minimise overlap (Doxey et al., 2017), reducing the degree of distortion and contamination, thereby improving the quality of the discourse representations: Hanert et al. (2017) found that a period of overnight sleep (vs. daytime wakefulness) increased a participant's ability to differentiate between an encoded image and a highly similar image, suggesting the possibility that hippocampal replay during sleep may have enhanced computations of pattern separation. In our Experiments 2 and 3, participants were exposed to six and four individual stories, respectively. Over time, these story representations might interfere with each other, increasing the chance of distortion. Assuming Hanert et al.'s (2017) finding extends to the verbal domain, it is possible that sleep consolidation may have helped reduce distortion by minimising the degree of overlap between stories as well as with any other unrelated information. Therefore, if sleep has an active role to play in maintaining the quality of discourse memories, it may do so via the strengthening of these memories and/or the separation of these memories from unrelated information. Finally, while these accounts prescribe an active role of sleep, it is also possible that given there is limited sensory input during sleep, discourse memory might be protected against external interference, resulting in less distortion (Hulme & Rodd, 2023; Paller et al., 2021). At present, it is unclear if the lower degree of distortion observed in the Sleep group is related to sleep actively consolidating discourse memory or passively protecting these memories from interference. Future work using polysomnography (PSG) or incorporating longer delay (e.g., 24 h) is needed to help tease these mechanisms apart.

Methodological contributions

In accessing discourse memory, most empirical studies (e.g., Aly & Moscovitch, 2010; Cohn-Sheehy et al., 2022; Reid et al., 2022; Wagner et al., 2001) focused on what was correctly retrieved and discarded recalled elements that were deemed to be distortions or errors. This is not an issue *per se*; nevertheless, in order to fully understand how discourse memory changes over time and sleep, it is important to examine these "unwanted" elements in a systematic way, because distortion is perhaps an inevitable characteristic of human memory (Bartlett, 1932). The fill-in-the-blank task developed for Experiments 2 and 3 provided us with a simple paradigm to examine responses that might be typically discarded, allowing us to explore in detail the nature of the 'errors' and to infer the underlying memory representations that are driving the responses. We are confident that with further refinement, the fill-in-the-blank task has the potential to be a widely adopted paradigm.

In addition to the paradigm itself, the task also highlighted the utility of LSA in assessing discourse memory. In the existing literature, story recall is usually quantified using subjective scoring protocols. For instance, in a recent study involving story recall, Denis, Dipierto, et al. (2021) had two human raters assign each recalled proposition into one of seven categories, including, for example, inference and importation. The key issue with this kind of scoring approach is that the boundary between each category is not always clear-cut, making it difficult for

future studies to replicate.[5] Therefore, there is a clear need for more objective scoring protocols. This prompted us to explore in Experiment 2 the utility of LSA. Complementing the subjective categorisation approach, our LSA metric showed that after a sleep opportunity (vs. wakefulness), participants' responses were closer to the target words in LSA semantic space, suggesting a lower degree of distortion. This LSA measure was then validated in the confirmatory replication in Experiment 3, yielding essentially the same results. Note, however, that this LSA measure was not intended to replace our categorisation metric, as what they captured are somewhat different, albeit related. For instance, our LSA measure cannot take a story's event model into account, so while antonyms like *increase* and *decrease* are high in LSA-cosine (0.82), one of them is likely to contradict a story's event model. Therefore, future work is required to refine and modify the LSA measure before it can serve as a substitute to subjective approaches.

Limitations and outstanding questions

In the study phase across the three experiments, participants encoded a relatively large amount of linguistic materials (e.g., 40 word pairs and 4 fairly long stories in Experiments 1B and 3). Some existing studies have shown that high information loads can reduce or even eliminate the benefit of sleep in declarative memory (Feld et al., 2016; Kolibius et al., 2021). Therefore, it is possible that our high information load occluded a sleep effect in discourse memory. However, we think this is unlikely, because (1) there was a clear sleep benefit in our paired-associate learning task, despite the high information load, and (2) some prior studies had revealed sleep-related memory effects even when participants encoded a large amount of varied information and completed various outcome measures (e.g., Schönauer et al., 2014). In other words, a high information/test load may not always compromise the detection of sleep-related memory effects, but it certainly is a factor that warrants attention when interpreting our data.

Another limitation of our current study is that we did not differentiate between various elements of a naturalistic story, which contains elements such as temporal-spatial contexts, protagonists, goals, and actions. Some elements, such as causal chains, are more central to a story, and are known to be more memorable (Mandler & Johnson, 1977; Mandler et al., 1980; Nezworski et al., 1982; Omanson, 1982; Stein & Glenn, 1979; Trabasso & van den Broek, 1985). The experiments reported in this article did not distinguish between these elements, and since sleep may have differential effects on central vs. peripheral components of an episode (e.g., Payne et al., 2008; but see Cohn-Sheehy et al., 2022), future work is needed to investigate whether various discourse elements may be differentially affected by sleep.

Like any other studies using an AM-PM/PM-AM design (e.g., Morgan et al., 2019), our studies were naturally confounded with time-of-day. To test whether it has an effect on memory performance, we included AM/PM control groups in Experiment 1 and took baseline measures immediately after study in Experiments 2 and 3. Performance on free word/story recall and paired-associate learning was equivalent between morning and evening, but both sentence recognition (Experiment 1) and fill-in-the-blank tasks (Experiment 2) revealed time-of-day influences, consistent with prior findings that reading strategy varies between morning and evening (Lorenzetti & Natale, 1996; Oakhill, 1986). This suggests that any difference between the Wake and Sleep groups after the 12-hr delay may be partially related to differences in reading strategy in the morning and evening. At present, our best evidence against this possibility comes from Experiment 3, where the Wake and Sleep groups were well-matched in terms of baseline performance in fill-in-the-blank, and still, there were clear between-group differences 12 h later. Having said that, future nap studies are needed to eliminate any time-of-day confounds (e.g., Horváth et al., 2015; Shaw & Monaghan, 2017).

Our study only tracked discourse memory over 12 h. Should sleep exhibit enduring effects on discourse memory with tangible implications for cognition, the advantages we noted ought to persist beyond the initial day (e.g., Lutz et al., 2017). Thus, future research endeavours could explore the trajectory of discourse memory over extended durations.

Finally, our studies were strictly behavioural, so it is not possible to draw any conclusion about the exact neurocognitive mechanism(s) that drive the changes in discourse representation. For instance, our data do not allow us to readily tease apart an active and a passive account for the effect of sleep. Studies using a nap design and with PSG will help address this question, and so will a study with a 24-hour interval between Session 1 and 2 (e.g., Gaskell et al., 2019; Experiment 2).

Conclusion

The episodic context account (Gaskell et al., 2019) predicts that at the point of language comprehension, episodic memory would come into play by binding different elements (e.g., words & concepts) together to form a discourse representation that is relatively abstract in nature. This representation is assumed to guide comprehension on-line, potentially by facilitating discourse retention and the formation of event models. Furthermore, just like any other newly formed episodic memory, these discourse representations are predicted to be susceptible to sleep-related effects. In testing this prediction, we used both recognition- and recall-based procedures to index discourse memory, representing arguably the most comprehensive examination to-date of how overnight sleep (vs. daytime wakefulness) may influence discourse memory that is schema-consistent. In both sentence recognition and free story recall, we found no evidence of sleep exerting an influence on discourse memory. In cued recall (i.e., fill-in-the-blank), however, we found that the degree of time-related distortion was lower after sleep than after wake, regardless of whether this was measured categorically or continuously, in two separate experiments ($N = 192$). Overall, these findings suggest that the effect of sleep on discourse memory is relatively modest and reject a strong version of the episodic context account, which needs to reconsider the centrality of sleep in discourse maintenance. Instead, we argue that findings from cued recall support a nuanced episodic context account such that the effect of sleep on discourse memory may [1] be constrained by the retrieval processes (recall vs. recognition; item vs. associative) and [2] primarily lie on a qualitative rather than a quantitative level, especially when the stories being used are schema-consistent. Furthermore, we argued that a sleep effect may situate at the textbase level of the tripartite model of discourse processing, with the reason being that this level is perhaps neither too weak (surface) nor too strong (event model) for sleep to act on. Our research represents an important step in reconciling the existing inconsistency in the sleep and discourse memory literature and in understanding the contribution of declarative memory to day-to-day language comprehension. We suggest that episodic memory is one part of an array of mechanisms that together support the retention of discourse memory across time spent awake and asleep. There may be a particular value to sleep in terms of linking associated concepts in discourse memory.

CRediT authorship contribution statement

Matthew H.C. Mak: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Adam J. Curtis:** Investigation, Formal analysis. **Jennifer M. Rodd:** Conceptualization, Funding acquisition, Writing – review & editing. **M. Gareth Gaskell:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

[5] We acknowledge that our categorisation approach also suffers from the same problem.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the materials, data, and R scripts are publicly available at https://osf.io/rk8j9/.

Acknowledgements

This research was supported by an Economic and Social Research Council (UK) grant (ES/T008571/1) to Gareth Gaskell and Jennifer Rodd. We thank Alice O'Hagan for her research assistance, Aidan Horner for his input on the conceptualisation stage for Experiment 2, and Jerry Fisher for sharing the materials and data from Fisher and Radvansky (2018).

Appendix A

A.1. Lexical properties of the 180 words used in word recall (Experiment 1A)

	Mean	SD	Min	Max
N of syllables	1.15	0.36	1	2
N of letters	4.64	0.96	3	7
Log Frequency	8.79	0.89	5.20	10.73
Concreteness	3.92	0.84	1.19	5

A.1. One of the nine word lists used in free word recall (Experiment 1A)

ache	click	guest	pine	spike
angle	coffin	gust	plank	sprint
arc	dumb	jeans	ruin	steel
badge	fail	jewel	scare	stiff

Appendix B

Word pairs in paired-associate learning (Experiments 1B and 3).

activity - fun
alley - street
almond - nut
atmosphere - air
axon - neuron
bacteria - fungus
bag - lunch
beach - sand
blender - food
blouse - shirt
brawl - bar
cabin - log
canyon - valley
chimney - smoke
cobbler - peach
compliment - thanks
creature - monster
custom - tradition
cut - blood
data - computer
difficulty - trouble
estimate - cost
fitness - exercise
flesh - skin
focus - camera
fog - mist
form - shape
fuzz - lint
grief - sorrow
holder - cup
hut - straw
mystery - novel
painter - artist
policy - rule
print - type
seam - stitch
topping - fudge
traitor - liar
trophy - prize
weather - climate

Appendix C

C.1. Sample story: Identification in CIA (Experiment 1)

One of the great espionage problems is the search for a reliable way of determining a person's identity. This is particularly important for agents working abroad. These agents were often forced to make contacts using only sketchy information. Several disasters abroad were caused by poor identification. For example, four years ago, several undercover agents died when they thought that the people they were meeting were their contacts, when in fact they were agents working for the other side. Lead agent, Linda Gill, was shot first. She died within minutes, exposing the mission. Before the rest of the group could react to the obvious danger, two more agents, Max Eagle and James Romney, took a bullet and went down like stones. This prompted the CIA in Washington to create a Board of Identification. This was at the end of Nicolas Elder's term as Agency Head. The Board was

empowered to award twenty-thousand dollars to the first person who developed a method of determining identity accurately ninety-nine percent of the time for a wide variety of people. There had been a number of attempts to solve this problem. One early idea was to have fingerprints taken at predetermined meeting sites. These sites would be strategically located across the world. A match could be made between a fingerprint and a stored file. The similarity between the two could be used to determine identity. Later, some engineers approached the identity challenge. They considered a retinal scan method. One year, Les Busby discovered that each retina had a different pattern that varied from person to person. Busby reasoned that this could be used as an identification method. This idea was based on the variations in peoples' retinas. These patterns would be distinctive no matter where a person was from. Busby even devised a special retinal scan helmet for people to wear. This method of determining identification captured the imagination of many the agency's administrators. Among those administrators were Cassell, Haynes, Hartley, and Nelson. A final idea was to use the DNA-based computer imaging system. A DNA imaging system is a device of great accuracy that can be used in most everyday situations. Early chemical- and spectral-based DNA identification methods were too cumbersome to be used abroad due to environmental changes. John Harrison was a self-taught computer game programmer. Early last year, Harrison invented and constructed four practical DNA identification systems. He completed his first system in April and submitted it to the Identification Board, but was turned down. The initial test of one of Harrison's systems was made in June. This was done abroad at a diplomatic conference. This first test of a DNA-based system was a grand success. He then built three more instruments, each smaller and more accurate than its predecessor. In August, Harrison's fourth system was tested on a trip to Egypt. It was found to be in error for only one person in a thousand. Although his systems all met the standards set up by the Board of Identification, he was not awarded any money until November, when he received five thousand dollars. A prominent member of the Board was Phil Marks. He was more impressed by the engineers. Marks thought that the programmer's device was less reliable than the work of the 'real' scientists. After several months, Harrison was taken under the wing of Senator Morris. Harrison ultimately claimed his reward money the following year. The newer DNA image identification systems are, broadly speaking, small, light-weight devices. A DNA sampling tube is hidden in a purse, briefcase, or clothing. As such, it remains available wherever the agent travels. The recent identification systems may be accurate to within one in ten thousand people.

C.2. Sample probes

Verbatim: Several disasters were caused by poor identification.
Paraphrase: Identity could be determined by getting the similarity between the two.
Inference: Harrison was a highly skilled programmer and brilliant inventor.Wrong: Marks himself was an amateur engineer.

C.3. Abridged version in Experiments 2 and 3

One of the great espionage problems is the search for a reliable way of determining a person's identity. This is particularly important for agents working abroad. These agents were often forced to make contacts using only sketchy information. Several disasters abroad were caused by poor identification. For example, four years ago, several undercover agents died when they thought that the people they were meeting were their contacts, when in fact they were agents working for the other side. This prompted the CIA in Washington to create a Board of Identification. The Board was empowered to award twenty-thousand dollars to the first person who developed a method of determining identity accurately ninety-nine percent of the time. There had been a number of attempts to solve this problem. One early idea was to have fingerprints taken at predetermined meeting sites. These sites would be strategically located across the world. A match could be made between a fingerprint and a stored file. Later, some engineers approached the identity challenge. One year, Les Busby discovered that each retina had a different pattern that varied from person to person. Busby reasoned that this could be used as an identification method. This idea was based on the variation in peoples' retinas. These patterns would be distinctive no matter where a person was from. Busby even devised a special retinal scan helmet for people to wear. Another idea was to use the DNA-based computer imaging system. A DNA imaging system is a device of great accuracy that can be used in most everyday situations. Early chemical- and spectral-based DNA identification methods were too cumbersome to be used abroad. John Harrison was a self-taught computer game programmer. Early last year, Harrison patented four practical DNA identification systems. He completed his first system in April and submitted it to the Identification Board, but was turned down. The initial test of one of Harrison's systems was made in June. This was done abroad at a diplomatic conference. This first test of a DNA-based system was a success. He then built three more instruments, each smaller and more accurate than its predecessor. In August, Harrison's fourth system was tested on a trip to Egypt. It was found to be in error for only one person in a thousand. Although his systems all met the standards set up by the Board of Identification, he was not awarded any money until November, when he received five thousand dollars. A prominent member of the Board was Phil Marks. He was more impressed by the engineers. Marks thought that the programmer's device was less reliable than the work of 'real' scientists. After several months, Harrison was taken under the wing of a senator. Harrison ultimately claimed his reward money the following year. The newer DNA image identification systems are, broadly speaking, light-weight devices. A DNA sampling tube is hidden in a purse, briefcase, or clothing. As such, it remains available wherever the agent travels.

Appendix D

Target words and the accompanying phrase/sentence in the fill-in-the-blank task (Exp 2 and 3).

Story	*Target word*	*Phrase/Sentenc*
Beanie	*chew*	*worth several thousand dollars for a/an_____toy.*
Beanie	*craze*	*then returned during the_____sometimes made mistakes.*
Beanie	*cuddly*	*because they are_____little toys that come,*
Beanie	*fell*	*The price of beanie babies_____, and very,*
Beanie	*fetched*	*buying rare beanie babies that_____high prices.*
Beanie	*high-quality*	*children weren't playing with the most_____toys.*
Beanie	*irresistible*	*wealthy people would find beanie babies_____,*
Beanie	*limited-edition*	*sixty-thousand dollars was paid for one_____beanie…*
Beanie	*matter*	*most of Lakewood was involved in the_____.*

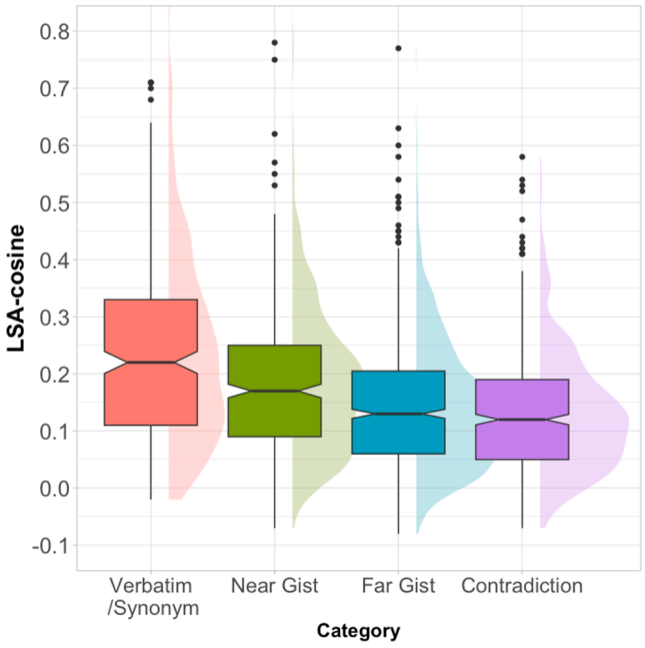
(*continued on next page*)

(*continued*)

Story	*Target word*	*Phrase/Sentenc*
Beanie	*neglected*	*Ordinary business was being_____throughout the city.*
Beanie	*neighbours*	*Almost everyone tried to outdo their_____,*
Beanie	*panic*	*wanted anymore. This realisation led to a/an_____.*
Beanie	*popular*	*frogs, alligators, and the_____parrot beanie baby.*
Beanie	*profitable*	*a/an_____bar was exchanged for one hippo,*
Beanie	*uncool*	*did not collect beanie babies were deemed_____.*
Beanie	*welcome*	*bright cheery beanie baby is a/an_____sight.*
CIA	*approached*	*Later, some engineers_____the identity challenge. One*
CIA	*cumbersome*	*methods were too_____to be used abroad.*
CIA	*diplomatic*	*This was done abroad at a/an_____conference.*
CIA	*disasters*	*Several_____abroad were caused by poor identification.*
CIA	*distinctive*	*These patterns would be_____no matter where,*
CIA	*empowered*	*The Board was_____to award twenty-thousand dollars,*
CIA	*everyday*	*that can be used in most_____situations.*
CIA	*instruments*	*He then built three more_____, each smaller,*
CIA	*light-weight*	*identification systems are, broadly speaking,_____devices.*
CIA	*patented*	*year, Harrison_____four practical DNA identification systems.*
CIA	*predetermined*	*to have fingerprints taken at_____meeting sites.*
CIA	*prominent*	*A/An_____member of the Board was Phil*
CIA	*reliable*	*less_____than the work of 'real' scientists*
CIA	*self-taught*	*John Harrison was a/an_____computer game programmer.*
CIA	*sketchy*	*forced to make contacts using only_____information.*
CIA	*variation*	*was based on the_____in peoples' retinas.*
Farmer	*anxious*	*Tess was particularly_____. He wanted to warn,*
Farmer	*detailed*	*the_____confessions that have been made public,*
Farmer	*episode*	*rebellion was an important_____in our town,*
Farmer	*extremist*	*hired the services of an anti-government_____, George*
Farmer	*grievances*	*It stemmed from a set of long-standing_____.*
Farmer	*intelligent*	*not clear why a/an_____pig farmer like,*
Farmer	*intensified*	*November. The plot_____government suspicions of farmers.*
Farmer	*invented*	*the whole story was_____by Steve Flett,*
Farmer	*mysteries*	*There are many_____about the farmers' rebellion,*
Farmer	*odd*	*Also it is_____that although the letter,*
Farmer	*pro-farmer*	*explosion might kill friendly_____members of the,*
Farmer	*rigorous*	*led to the_____enforcement of the Smythe,*
Farmer	*scheme*	*thought that such a/an _____would work or,*
Farmer	*severe*	*farmers had been subjected to_____environmental laws.*
Farmer	*strengthen*	*so that he could_____his position in*
Farmer	*wealthy*	*Collins enlisted his_____cousin Billy Hawkins in*
NY	*alienated*	*Black_____the Drug Corps by accusing the*
NY	*bullying*	*Later, Stevens succeeded in_____the city council*
NY	*confusion*	*The trial broke up in_____and Stevens was,*
NY	*conspiracy*	*Black charged them all with_____, which,*
NY	*detested*	*Black_____the direction the city was taking.*
NY	*endurance*	*deal in high_____technologies that were in,*
NY	*expansion*	*The_____that took place would not have,*
NY	*financial*	*future of the city lay in_____management.*
NY	*invaded*	*The Drug Corps_____Black's house on January,*
NY	*overcrowded*	*The offices in New York are very_____.*
NY	*provoked*	*This_____Major George Johnston and he declared,*
NY	*reputable*	*Stevens had other, less_____, business ventures. Stevens,*
NY	*stifle*	*Black wanted to_____the drug traffic and,*
NY	*tempting*	*potential of this growing business were very_____.*
NY	*unchecked*	*They ran the city_____by the state…*
NY	*unreliable*	*that sources of technology from Vermont were_____.*

Appendix E

Distribution of LSA-cosine for each unique target-response pair across the four categories in Experiment 2. The notch displays a confidence interval around the median which is based on median +/−1.58 * IQR/sqrt(n). The density plots represent smoothed density.



Appendix F

P values of Sleep-Wake comparison with and without baseline performance as a covariate across tasks in Experiments 2 and 3.

Wake vs. Sleep comparison	P value (With baseline as a covariate)	P value (Without baseline as a covariate)
Experiment 2		
Short story free recall (Second-New)	.674	.800
Short story free recall (Second-Repeated)	.864	.725
Verbatim	.131	.875
Verbatim/Synonym	.110	.729
Near Gist	.035*	.383
Far Gist	.042*	.508
Contradiction	.431	.810
LSA	.010*	.008*
Experiment 3		
Paired-associate	<.001	.189
Verbatim	.567	.799
Verbatim/Synonym	.020*	.147
Near Gist	.150	.097
Far Gist	.001*	.002*
Contradiction	.537	.661
LSA	.002*	.003*

Dropping baseline as a covariate increased the *p* values in all but one comparison and resulted in the sleep-wake comparison becoming non-significant in paired-associate learning, Near and Far Gist in Experiment 2, and Verbatim/Synonym in Experiment 3. These findings align with previous meta-analyses (e.g., Berres & Erdfelder, 2021; Lipinska et al., 2019) showing that a sleep effect is more likely to be detected when baseline performance is controlled for, as individual differences in memory encoding and retrieval may obscure any sleep-wake difference. Interestingly, the LSA analyses across the two experiments showed consistent results regardless of whether baseline performance was included as a covariate. Potentially, this may be due to the fact that the dependent variable (LSA-cosine) has a much wider score range (−0.08 to 0.95) than that in paired-associate learning (0 to 40) or categorisation approach (0 to 32), and this greater variability may have improved its sensitivity.

References

Abel, M., & Bäuml, K.-H. T. (2013). Adaptive memory: The influence of sleep and wake delay on the survival-processing effect. *Journal of Cognitive Psychology, 25*(8), 917–924. https://doi.org/10.1080/20445911.2013.825621

Abel, M., Haller, V., Köck, H., Pötschke, S., Heib, D., Schabus, M., & Bäuml, K.-H.-T. (2019). Sleep reduces the testing effect—But not after corrective feedback and prolonged retention interval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 272–287. https://doi.org/10.1037/xlm0000576

Adan, A., & Almirall, H. (1991). Horne & Östberg morningness-eveningness questionnaire: A reduced scale. *Personality and Individual Differences, 12*(3), 241–253. https://doi.org/10.1016/0191-8869(91)90110-W

Altmann, G. T. M., & Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. *Psychological Review, 126*(6), 817–840. https://doi.org/10.1037/rev0000154

Aly, M., & Moscovitch, M. (2010). The effects of sleep on episodic memory in older and younger adults. *Memory, 18*(3), 327–334. https://doi.org/10.1080/09658211003601548

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences, 21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Antony, J. W., & Paller, K. A. (2018). Retrieval and sleep both counteract the forgetting of spatial information. *Learning & Memory, 25*(6), 258–263. https://doi.org/10.1101/lm.046268.117

Ashton, J. E., & Cairney, S. A. (2021). Future-relevant memories are not selectively strengthened during sleep. *PLoS ONE, 16*. https://doi.org/10.1371/journal.pone.0258110

Baddeley, A., Vargha-Khadem, F., & Mishkin, M. (2001). Preserved recognition in a case of developmental amnesia: Implications for the acquisition of semantic memory? *Journal of Cognitive Neuroscience, 13*(3), 357–369.

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73*, 116–130. https://doi.org/10.1016/j.jml.2014.03.002

Ball, L., Mak, M. H. C., Ryskin, R. A., Curtis, A. J., Rodd, J. M., & Gaskell, M. G. (2024). The contribution of learning and memory processes to verb-specific syntactic processing. *PsyArix*. https://doi.org/10.31234/osf.io/4xpdz

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology* (p. xix, 317). Cambridge University Press.

Bastin, C., der Linden, M. V., Charnallet, A., Denby, C., Montaldi, D., Roberts, N., & Andrew, M. R. (2004). Dissociation between recall and recognition memory performance in an amnesic patient with hippocampal damage following carbon monoxide poisoning. *Neurocase, 10*(4), 330–344. https://doi.org/10.1080/13554790490507650

Bäuml, K. H. T., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect: It enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of Experimental Psychology: Learning Memory and Cognition, 40*(6), 1568–1581. https://doi.org/10.1037/xlm0000025

Bayley, P. J., O'Reilly, R. C., Curran, T., & Squire, L. R. (2008). New semantic learning in patients with large medial temporal lobe lesions. *Hippocampus, 18*(6), 575–583. https://doi.org/10.1002/hipo.20417

Berres, S., & Erdfelder, E. (2021). The sleep benefit in episodic memory: An integrative review and a meta-analysis. *Psychological Bulletin, 147*, 1309–1353. https://doi.org/10.1037/bul0000350

Blagrove, M., & Akehurst, L. (2000). Effects of sleep loss on confidence-accuracy relationships for reasoning and eyewitness memory. *Journal of Experimental Psychology: Applied, 6*(1), 59–73. https://doi.org/10.1037/0278-7393.6.1.59

Blank, I., Duff, M., Brown-Schmidt, S., & Fedorenko, E. (2016). Expanding the language network: Domain-specific hippocampal recruitment during high-level linguistic processing. *BioRxiv*. , Article 091900. https://doi.org/10.1101/091900

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological Research, 76*(2), 192–203. https://doi.org/10.1007/s00426-011-0335-6

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*, 396–439. https://doi.org/10.1016/j.dr.2004.08.005

Cohn-Sheehy, B. I., Delarazan, A. I., Crivelli-Decker, J. E., Reagh, Z. M., Mundada, N. S., Yonelinas, A. P., Zacks, J. M., & Ranganath, C. (2022). Narratives bridge the divide between distant events in episodic memory. *Memory & Cognition, 50*(3), 478–494. https://doi.org/10.3758/s13421-021-01178-x

Curtis, A. J., Mak, M. H. C., Chen, S., Rodd, J. M., & Gaskell, M. G. (2022). Word-meaning priming extends beyond homonyms. *Cognition, 226*, Article 105175. https://doi.org/10.1016/j.cognition.2022.105175

Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences, 100*(4), 2157–2162. https://doi.org/10.1073/pnas.0337195100

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., Dipierto, C., Spreng, N., Schacter, D., Stickgold, R., & Payne, J. (2021). Sleep and testing both strengthen and distort story recollection. *PsyArXiv*. https://doi.org/10.31234/osf.io/abt56

Denis, D., Mylonas, D., Poskanzer, C., Bursal, V., Payne, J. D., & Stickgold, R. (2021). Sleep spindles preferentially consolidate weakly encoded memories. *Journal of Neuroscience, 41*(18), 4088–4099. https://doi.org/10.1523/JNEUROSCI.0818-20.2021

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

Doolen, A. C., & Radvansky, G. A. (2021). A novel study: Long-lasting event memory. *Memory, 29*(8), 963–982. https://doi.org/10.1080/09658211.2021.1953079

Doxey, C. R., Hodges, C. B., Bodily, T. A., Muncy, N. M., & Kirwan, C. B. (2017). The effects of sleep on the neural correlates of pattern separation. *Hippocampus, 28*(2), 108–120. https://doi.org/10.1002/hipo.22814

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General, 136*, 169–183. https://doi.org/10.1037/0096-3445.136.2.169

Drosopoulos, S., Wagner, U., & Born, J. (2005). Sleep enhances explicit recollection in recognition memory. *Learning & Memory, 12*(1), 44–51. https://doi.org/10.1101/lm.83805

Duff, M. C., & Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. *Frontiers in Human Neuroscience, 6*. https://doi.org/10.3389/fnhum.2012.00069

Duff, M. C., & Brown-Schmidt, S. (2017). Hippocampal Contributions to Language Use and Processing. In D. E. Hannula, & M. C. Duff (Eds.), *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition* (pp. 503–536). Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_16.

Duff, M. C., Covington, N. V., Hilverman, C., & Cohen, N. J. (2020). Semantic memory and the hippocampus: Revisiting, reaffirming, and extending the reach of their critical relationship. *Frontiers in Human Neuroscience, 13*(January), 1–17. https://doi.org/10.3389/fnhum.2019.00471

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Empson, J. A. C., Hearne, K. M. T., & Tilley, A. J. (1981). REM sleep and remeniscence. In W. P. Koella (Ed.), *Sleep* (pp. 364–366). Basel: Karger.

Feld, G. B., Weis, P. P., & Born, J. (2016). The Limited Capacity of Sleep-Dependent Memory Consolidation. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01368

Fisher, J. S., & Radvansky, G. A. (2018). Patterns of forgetting. *Journal of Memory and Language, 102*, 130–141. https://doi.org/10.1016/j.jml.2018.05.008

Fletcher, C. R., & Chrysler, S. T. (1990). Surface forms, textbases, and situation models: Recognition memory for three types of textual information. *Discourse Processes, 13*, 175–190. https://doi.org/10.1080/01638539009544752

Friedrich, M., Wilhelm, I., Mölle, M., Born, J., & Friederici, A. D. (2017). The sleeping infant brain anticipates development. *Current Biology, 27*(15), 2374–2380.e3. https://doi.org/10.1016/j.cub.2017.06.070

Gaskell, M. G., Cairney, S. A., & Rodd, J. M. (2019). Contextual priming of word meanings is stabilized over sleep. *Cognition, 182*, 109–126. https://doi.org/10.1016/j.cognition.2018.09.007

Girardeau, G., Inema, I., & Buzsáki, G. (2017). Reactivations of emotional memory in the hippocampus-amygdala system during sleep. *Nature Neuroscience, 20*(11), 1634–1642. https://doi.org/10.1038/nn.4637

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163–189. https://doi.org/10.1146/annurev.psych.48.1.163

Hanert, A., Weber, F. D., Pedersen, A., Born, J., & Bartsch, T. (2017). Sleep in Humans Stabilizes Pattern Separation Performance. *Journal of Neuroscience, 37*(50), 12238–12246. https://doi.org/10.1523/JNEUROSCI.1189-17.2017

Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Holdstock, J. S., Mayes, A. R., Gong, Q. Y., Roberts, N., & Kapur, N. (2005). Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. *Hippocampus, 15*(2), 203–215. https://doi.org/10.1002/hipo.20046

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

Horváth, K., Myers, K., Foster, R., & Plunkett, K. (2015). Napping facilitates word learning in early lexical development. *Journal of Sleep Research, 24*(5), 503–509. https://doi.org/10.1111/jsr.12306

Hulme, R. C., & Rodd, J. M. (2023). The Role of Sleep in Learning New Meanings for Familiar Words through Stories. *Journal of Cognition, 6*(1), 27. https://doi.org/10.5334/joc.282

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General, 122*(2), 139–154. https://doi.org/10.1037/0096-3445.122.2.139

James, E., Gaskell, M. G., & Henderson, L. M. (2020). Sleep-dependent consolidation in children with comprehension and vocabulary weaknesses: It'll be alright on the night? *Journal of Child Psychology and Psychiatry, 61*(10), 1104–1115. https://doi.org/10.1111/jcpp.13253

Kintsch, W. (1992). A cognitive architecture for comprehension. In *Cognition: Conceptual and methodological issues* (pp. 143–163). American Psychological Association. https://doi.org/10.1037/10564-006.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*(2), 133–159. https://doi.org/10.1016/0749-596X(90)90069-C

Kolibius, L. D., Born, J., & Feld, G. B. (2021). Vast Amounts of Encoded Items Nullify but Do Not Reverse the Effect of Sleep on Declarative Memory. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.607070

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research, 17*(1), 3–10. https://doi.org/10.1111/j.1365-2869.2008.00622.x

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course* (p. xiii, 264). Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lipinska, G., Stuart, B., Thomas, K. G. F., Baldwin, D. S., & Bolinger, E. (2019). Preferential consolidation of emotional memory during sleep: A meta-analysis. *Frontiers in Psychology, 10*, Article 1014. https://doi.org/10.3389/fpsyg.2019.01014

Lo, J. C., Dijk, D.-J., & Groeger, J. A. (2014). Comparing the effects of nocturnal sleep and daytime napping on declarative memory consolidation. *PLoS ONE, 9*(9), Article e108100. https://doi.org/10.1371/journal.pone.0108100

Lorenzetti, R., & Natale, V. (1996). Time of day and processing strategies in narrative comprehension. *British Journal of Psychology, 87*(2), 209–221. https://doi.org/10.1111/j.2044-8295.1996.tb02586.x

Lutz, N. D., Diekelmann, S., Hinse-Stern, P., Born, J., & Rauss, K. (2017). Sleep supports the slow abstraction of GIST from visual perceptual memories. *Scientific Reports, 7* (October 2016), 1–9. https://doi.org/10.1038/srep42950

Mak, M. H. C. (2021). Children's motivation to learn at home during the COVID-19 pandemic: Insights from Indian parents. *Frontiers in Education, 6*(October), 1–7. https://doi.org/10.3389/feduc.2021.744686

Mak, M. H. C., Curtis, A. J., Rodd, J. M., & Gaskell, M. G. (2023). Episodic memory and sleep are involved in the maintenance of context-specific lexical information. *Journal of Experimental Psychology: General.*. https://doi.org/10.1037/xge0001435

Mak, M. H. C., & Gaskell, M. G. (2023). Effects of sleep and retrieval practice on verbal paired-associate learning across 12 and 24 hours. *PsyArXiv*. https://psyarxiv.com/phe5j.

Mak, M. H. C., Hsiao, Y., & Nation, K. (2021). Lexical connectivity effects in immediate serial recall of words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(12), 1971–1997. https://doi.org/10.1037/xlm0001089

Mak, M. H. C., O'Hagan, A., Horner, A. J., & Gaskell, M. G. (2023). A registered report testing the effect of sleep on Deese-Roediger-McDermott false memory: Greater lure and veridical recall but fewer intrusions after sleep. *Royal Society Open Science, 10* (12). https://doi.org/10.1098/rsos.220595

Mak, M. H. C., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin and Review, 27*(5), 1059–1069. https://doi.org/10.3758/s13423-020-01773-0

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*(1), 111–151. https://doi.org/10.1016/0010-0285(77)90006-8

Mandler, J. M., Scribner, S., Cole, M., & DeForest, M. (1980). Cross-Cultural Invariance in Story Recall. *Child Development, 51*(1), 19–26. https://doi.org/10.2307/1129585

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Hunkin, N. M., & Roberts, N. (2002). Relative sparing of item recognition memory in a patient with adult-onset damage limited to the hippocampus. *Hippocampus, 12*(3), 325–340. https://doi.org/10.1002/hipo.1111

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. https://doi.org/10.1037/0033-295X.102.3.419

Milivojevic, B., Varadinov, M., Vicente Grabovetsky, A., Collin, S. H. P., & Doeller, C. F. (2016). Coding of event nodes and narrative context in the hippocampus. *The Journal of Neuroscience, 36*(49), 12412–12424. https://doi.org/10.1523/JNEUROSCI.2889-15.2016

Miyamoto, D., Hirai, D., & Murayama, M. (2017). The roles of cortical slow waves in synaptic plasticity and memory consolidation. *Frontiers in Neural Circuits, 11*, Article 92. https://doi.org/10.3389/fncir.2017.00092

Morgan, D. P., Tamminen, J., Seale-Carlisle, T. M., & Mickes, L. (2019). The impact of sleep on eyewitness identifications. *Royal Society Open Science, 6*(12), Article 170501. https://doi.org/10.1098/rsos.170501

Natale, V., & Lorenzetti, R. (1997). Influences of morningness-eveningness and time of day on narrative comprehension. *Personality and Individual Differences, 23*(4), 685–690. https://doi.org/10.1016/S0191-8869(97)00059-7

Nation, K., & Snowling, M. J. (1998). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language, 39*(1), 85–101. https://doi.org/10.1006/jmla.1998.2564

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nezworski, T., Stein, N. L., & Trabasso, T. (1982). Story structure versus content in children's recall. *Journal of Verbal Learning and Verbal Behavior, 21*(2), 196–206. https://doi.org/10.1016/S0022-5371(82)90561-8

Oakhill, J. (1986). Effects of time of day on the integration of information in text. *British Journal of Psychology, 77*(4), 481–488. https://doi.org/10.1111/j.2044-8295.1986.tb02212.x

Omanson, R. C. (1982). The relation between Centrality and story category variation. *Journal of Verbal Learning and Verbal Behavior, 21*(3), 326–337. https://doi.org/10.1016/S0022-5371(82)90648-X

Paller, K. A., Creery, J. D., & Schechtman, E. (2021). Memory and sleep: How sleep cognition can change the waking mind for the better. *Annual Review of Psychology, 72* (1), 123–150. https://doi.org/10.1146/annurev-psych-010419-050815

Payne, J. D., Chambers, A. M., & Kensinger, E. A. (2012). Sleep promotes lasting changes in selective memory for emotional scenes. *Frontiers in Integrative Neuroscience, 6*, Article 108. https://doi.org/10.3389/fnint.2012.00108

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep Preferentially Enhances Memory for Emotional Components of Scenes. *Psychological Science, 19*(8), 781–788. https://doi.org/10.1111/j.1467-9280.2008.02157.x

Payne, J. D., Stickgold, R., Wamsley, E., Gibler, K., & Kensinger, E. A. (2010). A brief daytime nap selectively consolidates memory for emotional aspects of scenes. *Manuscript submitted for publication*.

Petros, T. V., Beckwith, B. E., & Anderson, M. (1990). Individual differences in the effects of time of day and passage difficulty on prose memory in adults. *British Journal of Psychology, 81*(1), 63–72. https://doi.org/10.1111/j.2044-8295.1990.tb02346.x

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

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9*(4), 534–547. https://doi.org/10.1162/jocn.1997.9.4.534

Radvansky, G. A. (2012). Across the event horizon. *Current Directions in Psychological Science, 21*, 269–272. https://doi.org/10.1177/0963721412451274

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging, 16*(1), 145–160. https://doi.org/10.1037/0882-7974.16.1.145

Rasch, B., & Born, J. (2007). Maintaining memories by reactivation. *Current Opinion in Neurobiology, 17*(6), 698–703. https://doi.org/10.1016/j.conb.2007.11.007

Reid, A., Bloxham, A., Carr, M., van Rijn, E., Basoudan, N., Tulip, C., & Blagrove, M. (2022). Effects of sleep on positive, negative and neutral valenced story and image memory. *British Journal of Psychology, 113*(3), 777–797. https://doi.org/10.1111/bjop.12559

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How Fuzzy-Trace Theory Predicts True and False Memories for Words, Sentences, and Narratives. *Journal of Applied Research in Memory and Cognition, 5*(1), 1–9. https://doi.org/10.1016/j.jarmac.2015.12.003

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language, 87*, 16–37. https://doi.org/10.1016/j.jml.2015.10.006

Rodd, J. M., Lopez Cutrin, B., Kirsch, H., Millar, A., & Davis, M. H. (2013). Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language, 68* (2), 180–198. https://doi.org/10.1016/j.jml.2012.08.002

Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition, 2* (1), 95–100. https://doi.org/10.3758/BF03197498

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*(3), 293–304. https://doi.org/10.1177/1745691614528214

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di linguista, 20*, 33–53.

Saletin, J. M., Goldstein, A. N., & Walker, M. P. (2011). The role of sleep in directed forgetting and remembering of human memories. *Cerebral Cortex, 21*(11), 2534–2541. https://doi.org/10.1093/cercor/bhr034

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), 1. https://doi.org/10.1038/s41598-017-12884-5

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25*, 279–294. https://doi.org/10.1016/0749-596X(86)90002-1

Schoch, S. F., Cordi, M. J., & Rasch, B. (2017). Modulating influences of memory strength and sensitivity of the retrieval test on the detectability of the sleep consolidation effect. *Neurobiology of Learning and Memory, 145*, 181–189. https://doi.org/10.1016/j.nlm.2017.10.009

Schönauer, M., Pawlizki, A., Köck, C., & Gais, S. (2014). Exploring the effect of sleep and reduced interference on different forms of declarative memory. *Sleep, 37*(12), 1995–2007. https://doi.org/10.5665/sleep.4258

Scullin, M. K. (2013). Sleep, memory, and aging: The link between slow-wave sleep and episodic memory changes from younger to older adults. *Psychology & Aging, 28*(1), 105–114. https://doi.org/10.1037/a0028830

Seger, B. T., Wannagat, W., & Nieding, G. (2021). Children's surface, textbase, and situation model representations of written and illustrated written narrative text. *Reading and Writing, 34*(6), 1415–1440. https://doi.org/10.1007/s11145-020-10118-1

Shaw, J. J., & Monaghan, P. (2017). Lateralised sleep spindles relate to false memory generation. *Neuropsychologia, 107*, 60–67. https://doi.org/10.1016/j.neuropsychologia.2017.11.002

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Squire, L. R., & Dede, A. J. O. (2015). Conscious and unconscious memory systems. *Cold Spring Harbor Perspectives in Biology, 7*(3), Article a021667. https://doi.org/10.1101/cshperspect.a021667

Squire, L. R., Wixted, J. T., & Clark, R. E. (2007). Recognition memory and the medial temporal lobe: A new perspective. *Nature Reviews. Neuroscience, 8*(11), 872–883. https://doi.org/10.1038/nrn2154

Stein, N., & Glenn, C. (1979). *An Analysis of Story Comprehension in Elementary School Children. New Directions in Discourse Processing* (p. 2).

Takashima, A., Wagensveld, B., van Turennout, M., Zwitserlood, P., Hagoort, P., & Verhoeven, L. (2014). Training-induced neural plasticity in visual-word decoding and the role of syllables. *Neuropsychologia, 61*, 299–314. https://doi.org/10.1016/j.neuropsychologia.2014.06.017

Tamminen, J., & Gaskell, M. G. (2013). Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *Quarterly Journal of Experimental Psychology, 66*(5), 1001–1025. https://doi.org/10.1080/17470218.2012.724694

Tilley, A. J., & Empson, J. A. C. (1978). REM sleep and memory consolidation. *Biological Psychology, 6*(4), 293–300. https://doi.org/10.1016/0301-0511(78)90031-5

Trabasso, T., & Van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24*(5), 612–630. https://doi.org/10.1016/0749-596X(85)90049-X

van Dijk, T. A., Kintsch, W., & Van Dijk, T. A. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

van Rijn, E., Carter, N., McMurtrie, H., Willner, P., & Blagrove, M. T. (2017). Sleep does not cause false memories on a story-based test of suggestibility. *Consciousness and Cognition, 52*(October 2016), 39–46. https://doi.org/10.1016/j.concog.2017.04.010

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277*(5324), 376–380. https://doi.org/10.1126/science.277.5324.376

Wagner, U., Gais, S., & Born, J. (2001). Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learning & Memory, 8*(2), 112–119. https://doi.org/10.1101/lm.36801

Wagner, U., Kashyap, N., Diekelmann, S., & Born, J. (2007). The impact of post-learning sleep vs. Wakefulness on recognition memory for faces with different facial expressions. *Neurobiology of Learning and Memory, 87*(4), 679–687. https://doi.org/10.1016/j.nlm.2007.01.004

Wang, H.-C., Savage, G., Gaskell, M. G., Paulin, T., Robidoux, S., & Castles, A. (2017). Bedding down new words: Sleep promotes the emergence of lexical competition in visual word recognition. *Psychonomic Bulletin & Review, 24*(4), 1186–1193. https://doi.org/10.3758/s13423-016-1182-7

Williams, S. E., & Horst, J. S. (2014). Goodnight book: Sleep consolidation improves word learning via storybooks. *Frontiers in Psychology, 5*, Article 184. https://doi.org/10.3389/fpsyg.2014.00184

Wills, A., Walsh, C., Sharpe, P., & Mitchell, C. (2020). More on Bayesian ANOVA. Retrieved from https://www.andywills.info/rminr/anova4.html.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus, 20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864

Yonelinas, A., Ranganath, C., Ekstrom, A., & Wiltgen, B. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews. Neuroscience, 20*(6), 364–375. https://doi.org/10.1038/s41583-019-0150-4

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 920–933. https://doi.org/10.1037/0278-7393.20.4.920

Zwaan, R. A., & Brown, C. M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*(3), 289–327. https://doi.org/10.1080/01638539609544960

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162