Sleep-associated consolidation in app-based language learning

Emma James (emma.james@york.ac.uk)
Department of Psychology, University of York, York, YO10 5DD, UK

Yolanda G. Koutraki (yolanda.koutraki@memrise.com)
Memrise Ltd, 3-5 Fashion St, London, E1 6PX, UK

Hannah Tickle (hannahpamelatickle@gmail.com)
Memrise Ltd, 3-5 Fashion St, London, E1 6PX, UK

Abstract

Neuro-cognitive models of word learning propose a role for sleep in consolidating new words, yet evidence for sleep-associated memory benefits outside of experimental contexts is scarce. This study compared wake- and sleep-associated memory changes in data from *Memrise*, a publicly available language-learning app. Memory for foreign words and phrases remained very high in accuracy across a 7-12 hour delay, and there were no differences in forgetting between wake and sleep. However, learners were quicker to arrive at the correct translation after a period of sleep compared to wake. This sleep-associated benefit was seen for words but not phrases, and could not be fully accounted for by circadian differences in completion time. As such, we demonstrate that the behavioural benefits of sleep on vocabulary can be observed in real-world language learning, and discuss the promise for combining small-scale lab studies with naturally occurring datasets to understand learning outcomes.

Keywords: vocabulary; learning; memory; consolidation

Introduction

When preparing for a trip abroad, an important task is to consider communication in the local language. You might try to learn some key vocabulary (*aeroporto*, *vino*, *formaggio*), as well as useful phrases to help you get by (*dov'e il bar più vicino? – where is the nearest bar?).* Experimental studies suggest that sleep-based processes play an important role in supporting new language learning, yet scope for capitalising upon sleep in real-world contexts is unclear. In this study, we examine whether sleep-associated memory benefits can be observed in naturally occurring data sourced from *Memrise*, a foreign language learning app.

The role of sleep in language learning is described by the Complementary Learning Systems model (Davis & Gaskell, 2009). According to this model, two neural systems support the acquisition of new vocabulary: a rapid-learning hippocampal system, and a slower-learning neocortical system that retains knowledge for the longer-term. The hippocampal system is able to create a new memory trace very quickly upon hearing a new spoken word, binding together information about its phonological form and meaning. To avoid interference with known words, the neocortical system requires a slower process of consolidation to embed the new word within existing vocabulary. This consolidation process can happen "offline"—without further input or rehearsal—and numerous studies support an active role for sleep. Specifically, the hippocampus is proposed to replay newly learned words during sleep, allowing the new connections to become strengthened within neocortically stored vocabulary (Davis et al., 2009).

This consolidation process is marked by a number of observable changes in new word memory. A key experimental approach is to train new words in the morning or the evening, and assess memory both immediately and ~12 hours later. Participants who learn in the evening—and therefore sleep before the 12-hour test—tend to show less forgetting of new vocabulary than participants who learn in the morning and remain awake during the intervening period (Gais, Lucas, & Born, 2006). Some studies have even identified sleep-associated *improvements* in the recall of new words, as well as increased competition with existing vocabulary that marks integration of the new words into existing networks (e.g., Dumay & Gaskell, 2007). These findings are of core theoretical relevance for both first and second language learning (Lindsay & Gaskell, 2010). Furthermore, they lead to exciting practical questions—such as whether learning can be timed to best capitalise upon sleep-associated mechanisms, or whether focusing on sleep problems may help to remediate learning difficulties.

However, a fundamental issue precedes these more applied questions: there is not yet evidence that the behavioural benefits of sleep-associated consolidation are observable outside of tightly controlled experimental settings. Three core aspects of laboratory research suggest that external validation cannot be assumed. First, participating in a sleep experiment constitutes a highly salient event for participants: they are typically required to learn large amounts of information late in the evening, in an unfamiliar setting. This salient information may be prioritised during subsequent consolidation (e.g., Wilhelm et al., 2011), leading to larger estimates of sleep benefits than occur when information is learned in an everyday context. Second, the motivation of learners presumably differs between laboratory experiments and natural settings, given that the former tend to involve tightly controlled stimuli with limited long-term use. There is some evidence that motivation may enhance consolidation of newly learned information (Fischer & Born, 2009; Studte, Bridger, & Mecklinger, 2017), yet it is also possible that high motivation and engagement could mask potential benefits of

sleep. Third, many experimental tasks minimise feedback after learning to examine "pure" processes of memory consolidation, whereas learners in the real-world receive feedback to maximise retention of new knowledge. It may be that the benefits of feedback are sufficient enough to render offline processes redundant, especially considering that weaker memory traces that are more susceptible to sleep-associated benefits (Diekelmann, Wilhelm, & Born, 2009).

In sum, while sleep's role in memory consolidation is well-established, our key question here is whether there are observable benefits of these processes in naturalistic language learning. A number of cognitive scientists have recently advocated the use of naturally occurring datasets for external validation of lab-based findings (e.g., Goldstone & Lupyan, 2016; Paxton & Griffiths, 2017), to complement data collected via tightly controlled experiments. Only one study to our knowledge has used naturally occurring data to examine memory consolidation: Stafford and Haasnoot (2017) looked at procedural learning performance in an online game, and did *not* find evidence for sleep-associated benefits. Establishing the scalability of lab-based findings is thus an essential first step if we are to consider sleep as a target for supporting language learning.

In this study, we tested whether overnight periods (assumed to contain sleep) are more beneficial for language memory than daytime periods (assumed to contain wake) in data from *Memrise*. *Memrise* is a foreign language-learning platform that enables users to acquire vocabulary and phrases in their chosen language(s), using principles of spaced repetition and repeated retrieval practice. We selected sequential trials of app use that spanned half-day periods, and that were timed to contain either sleep or wake for the majority of users. Using these data, we were able to test two fundamental hypotheses. First, whether sleep benefits memory for foreign *vocabulary* compared to an equivalent period of wake. On the basis of experimental studies demonstrating sleep-associated improvements for new word-forms, we anticipated that vocabulary tests were most likely to show benefits for sleep versus wake, and that these benefits would not be attributable to circadian differences in performance. Second, we tested whether sleep also benefits memory for foreign *phrases*. Memory for larger chunks of linguistic information has scarcely been examined in the laboratory, and these data afforded the opportunity to explore more applied benefits for language consolidation.

Method

Initial sample

Data from 7700 *Memrise* users were collected between July and November 2018. All users were accessing the app through Android devices, and paid for a subscription to the service during this period. The largest proportion of users were in Europe (54%), but users spanned six continents. The majority of users were using the app to learn a single language, but 30% of users were learning more than one language during this period. Across users, 17 different "target" languages were being learned (most common: English, 3253 users; least common: Danish, 27 users), from an array of different "source" languages (most common: English, 3039 users; least common: Norwegian, 34 users). In total, there were 137 different source-target language combinations. All users consent to being over 13 years old, but no further demographic information is collected by the platform.

Although the data cannot be made publicly available, R Markdown files documenting the data processing, analysis, and output are available on the Open Science Framework (https://osf.io/skuzd). Access to the materials can be gained via free trials of the app downloaded from Google Play or the App Store (see https://www.memrise.com/apps/).

Learning and Test Tasks

Within the app, users complete a range of different tasks to support their learning. After being presented with new words and phrases, they complete multiple-choice tasks from written and/or spoken cues, and typing tasks that require translating and typing the new material in the target language. Given the experimental evidence that production tasks are most sensitive to sleep-associated benefits (see Diekelmann et al., 2009, for a review), we focused the present analyses on the typing tasks. Two tasks allowed us to assess word knowledge and phrase knowledge separately.

To assess word knowledge, users are presented with a word in their source language, and are required to type the corresponding translation in the target language (Figure 1a). This task places demands on phonological and orthographic knowledge of the new word-forms, as well as their translation. Users are provided with a constrained number of letter options to use in their response, and the trial ends automatically upon the production of the correct answer. Thus, while users *can* submit an incorrect response, they are also able to edit their responses until the correct answer is achieved. As a result, the proportion of trials correct is very high (90%) with little scope for variation, but trial completion time can additionally be analysed as a marker of learning (*median* = 6.46 s; *IQR* = 6.75 s, *max* = 589025 s).

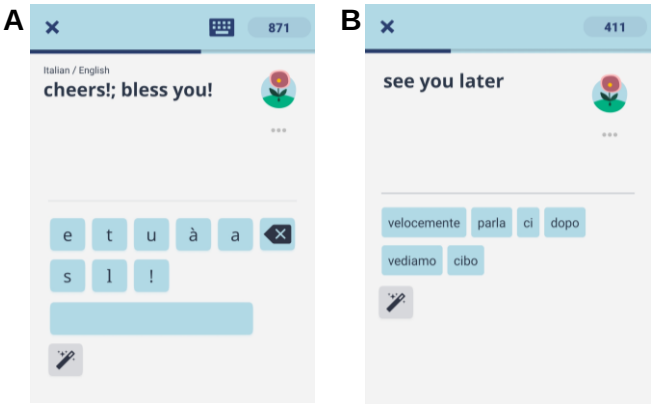

Figure 1: Example trials from a) the word task; and b) the phrase task.

The positive skew was reduced by excluding trials above 20 s (9.62%), under the assumption that trials beyond this likely reflect inattention rather than repeated attempts to arrive at the correct answer (new *median* = 5.94 s; *IQR* = 5.08 s).

To assess phrase knowledge, users are presented with a phrase in their source language, and are required to select and order the appropriate words from a constrained set of options to provide the translation (Figure 1b). This task places demands on understanding (but not producing) the correct vocabulary, and on syntax. The proportion of trials correct is very high (95%) given that users can make multiple attempts, and so completion time is additionally used as a marker of learning (*median* = 6.46 s; *IQR* = 6.49 s; *max* = 21504 s*).* As above, trials above 20 s were excluded (8.99%; new *median* = 5.98 s; *IQR* = 4.97 s).

Data Selection and Analyses

Data were processed using the *data.table* package (Dowle et al., 2019) in *R*. For each typing task (word, phrase), we selected trials that were preceded by the same test type for that item. For example, if the user was tested on *formaggio*, the trial was only included if the user's last test for *formaggio* was also a typing task, and not a multiple choice task. This enabled us to compute two dependent variables for analysis: the change in accuracy and change in completion time relative to this previous trial. Analysing change between trials (rather than raw trial performance) permitted the inclusion of multiple transitions per user/item where available.

For the main analyses, we were interested in changes in performance across half-day periods, assumed to contain either wake or sleep. Given that the timings of experimental studies comparing wake and sleep based periods show considerable variability, we followed the data selection protocol of Stafford and Haasnoot (2017) in their analysis of offline consolidation in a naturally occurring dataset. In line with their criteria, we selected trials that had between 7 and 12 hours since the previous test, with the present test being either between 5am-12pm (*sleep-associated changes*) or 5pm-12am (*wake-associated changes*). The word dataset contained 10,566 trials that met these criteria, spanning 1830 users and 115 target-source language combinations. The average timespan from the previous trial was 9.60 hours, and was approximately half an hour longer in the sleep (M = 9.85, SD = 1.44) compared to wake (M = 9.33, SD = 1.42) trials.[1] The phrase dataset contained 6856 trials that met these criteria, spanning 1514 users and 113 language combinations. The average timespan from the previous trial was 9.64 hours (SD = 1.42), and was again slightly longer in the sleep (M = 9.86, SD = 1.36) compared to wake (M = 9.40, SD = 1.44) trials in the phrase dataset. We included all trials that met the timing criteria, capturing broad variability in users' prior exposure to the items (lexical range: 2-493 tests; phrase range 2-186 tests). Thus, we examined sleep-associated processes across prolonged periods of consolidation, which are rarely examined in the lab.

To assess the effect of sleep on changes in accuracy, we fitted a cumulative link mixed model with sleep-wake category as a fixed effect and memory change (-1, 0, 1) as the dependent variable (package *ordinal*; Christensen (2015)). A memory change of -1 reflects an incorrect response to a previously correct trial, +1 reflections a correct response to a previously incorrect trial, and 0 can reflect either two sequentially correct or two sequentially incorrect trials. Both user ID and item (specific word or phrase) were initially entered as random intercepts, alongside the number of previous tests the user had experienced with that item. The latter was included to control for any differences in the stage of learning between conditions, and account for non-independence between tests of the same item for the same user. Indeed, its inclusion improved model fit in three of the four models (ps < .001; phrase accuracy p = .109), suggesting that it accounted for additional variance. Beyond this basic structure, we fitted the maximal model supported by the data: we used likelihood ratio tests to see whether any additional random intercepts improved the fit of the model (source language, target language), and if random slopes for the effect of sleep-wake category were warranted (threshold p < .2; Barr et al. (2013)).

To assess the effect of sleep on changes in completion time, we used a linear mixed model specified in the same way as above (package *lme4;* Bates et al. (2014)), and used *lmerTest* to compute statistical significance (Kuznetsova, Brockhoff, & Christensen, 2017). Details of all models are available on the OSF (https://osf.io/skuzd). Figures were made using the *ggplot2* package (Wickham, 2016).

Results

Offline Changes in Word Memory

Accuracy There was very little change in accuracy across the time period, with 89% of trials in each condition showing no change in performance relative to the trial before (i.e., either remaining correct, or remaining incorrect). The remaining trials showed greater forgetting than gains in performance. For the sleep condition, 9.26% of trials decreased in accuracy relative to the previous trial, compared to 1.90% showing an improvement in accuracy. For the wake condition, 9.84% of trials showed a decrease compared to 1.51% showing an improvement. This meant that overall, the proportion of trials correct decreased by 7.36% in the sleep condition and 8.33% in the wake condition (Figure 2a). This difference was not statistically significant (β = 0.09, SE = 0.11, Z = 0.82, p = .412).

Completion time We analysed changes in completion time for correct responses only (with no change in accuracy),

[1] Note that the pattern of results remained the same when analyses were re-run on a subset of data matched for timespan.

leaving 7697 trials in total (from 1598 users). There was a general increase in completion time over the 7-12 hour period, likely as the item was more accessible to memory after recent app use than when users returned after a longer break. Importantly, the magnitude of this increase differed between the sleep and wake conditions (β = -244.75, *SE* = 112.69, *t* = -2.17, *p* = .030; Figure 2b). Across a period of wake, the mean increase in completion time was 709 ms (*SD* = 4231 ms). Across sleep, this increase was smaller (*M* = 438 ms, *SD* = 4129 ms).

Offline Changes in Phrase Memory

Accuracy As with the word task, the majority of trials showed no change in accuracy relative to the trial before (sleep: 93.44%; wake: 92.23%). For the sleep condition, 5.32% percent of trials showed a decrease in memory accuracy (i.e., forgetting), whereas 1.24% showed an improvement. For the wake condition, 6.22% showed a decrease in accuracy, and 1.55% showed an improvement.

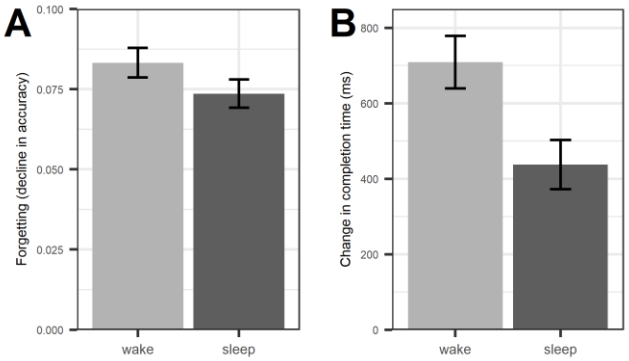

Figure 2: Wake- and sleep-associated changes in word task performance for a) Mean proportion forgotten; and b) Trial completion time. Lower values represent better memory retention in each graph. Error bars represent +/- 1 SE.

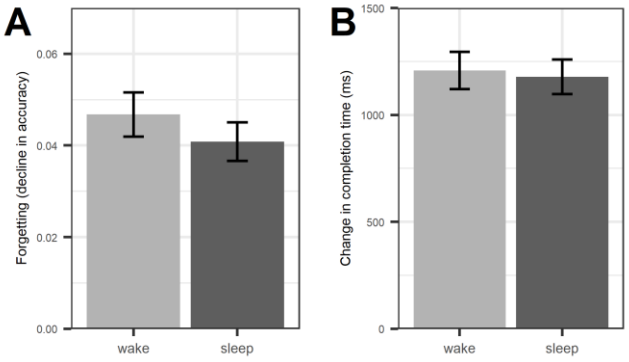

Figure 3: Wake- and sleep-associated changes in phrase task performance for a) Mean proportion forgotten; and b) Trial completion time. Lower values represent better memory retention in each graph. Error bars represent +/- 1 SE.

On average then, the proportion of trials correct decreased by 4.08% over assumed sleep, and 4.67% over assumed wake (Figure 3a). This difference was not statistically significant (β = 0.02, *SE* = 0.12, *Z* = 0.14, *p* = .889).

Completion time As above, only correct responses were modelled, leaving 5039 trials (from 1333 users). Completion time over the half-day period increased by a mean of 1178 ms (*SD* = 4226 ms) when it was assumed to contain sleep, and 1208 ms (*SD* = 4194 ms) when assumed to contain wake (Figure 3b). The difference between sleep and wake periods was not statistically significant (β = -44.27, *SE* = 130.67, *t* = -0.34, *p* = .735).

Consolidation or Fatigue?

The analyses revealed a possible benefit for sleep in memory for words, demonstrated by smaller increases in completion times over sleep relative to wake. We interpret this completion time benefit as users being able to more quickly retrieve the memory and/or arrive at the correct answer via repeated attempts. However, these analyses cannot elucidate the underlying mechanisms. From a Complementary Learning Systems perspective, sleep-associated changes in the memory representation are proposed to influence access to the new word. However, circadian influences may also affect completion time across these periods: if users are fatigued in the evening, they may be slower to complete the task than when they are refreshed in the morning, and vice versa for the wake condition. Although it is not possible to fully disentangle consolidation and circadian influences in these data, we examined this issue in two ways.

First, we tested for differences in performance at the initial test points in the selected word dataset. The above analyses were conducted on the *change* in performance across two trials spanning a 7-12 hour gap, whereas here we test whether performance differs in the first of those tests. If there are circadian differences in completion time, we would expect that performance would be slower in the sleep condition at the first test point (i.e., when users completed the activities in the evening) than in the wake condition (i.e., when the activities were completed in the morning). We re-ran the analysis using completion time at this first test point as the dependent variable, and time of day (morning, evening) as the fixed effect of interest. The data were subject to a Box-Cox transformation to ameliorate issues of non-normality, with median raw scores reported to aid interpretation. Completion times in the evening (*median* = 5.44 s; *IQR* = 4.24 s) were slightly slower than in the morning (*median* = 5.42 s; *IQR* = 4.31 s), but this difference was not statistically significant (β = 0.01, *SE* = 0.01, *t* = 0.95, *p* = .342).

To ensure that the lack of circadian influences are not an artefact of the data selection process in some way, our second approach was to examine differences in trial completion time between our two time periods of interest (5am-12pm; 5pm-12am) across the whole sample. We selected all trials for the word production task that took place during these two periods (regardless of the time/nature of the preceding trial), and

modelled differences in completion time using the same approach as above. There was a significant difference in completion time between the two time periods ($\beta = 0.01$, $SE = 0.003$, $t = 4.53$, $p < .001$): trials were completed slightly faster in the morning (*median* = 5.94 s, *IQR* = 5.04 s) than in the evening (*median* = 5.98 s, *IQR* = 5.09 s). However, the difference in completion time was extremely small (difference in medians = 45 ms; difference in means = 33 ms), reaching statistical significance in the present model with > 1.69 million trials. This result suggests that there are circadian influences on completion time, but that these unlikely fully account for the effects seen in the sleep-wake analysis (difference in mean change = 295 ms).

Discussion

We asked whether the behavioural benefits of sleep-associated consolidation can be observed in app-based language learning. In using a naturally occurring dataset from *Memrise*, wake- and sleep-associated memory changes could be examined across prolonged periods of memory consolidation, and when learners could freely engage with the activities at times to suit them. Whilst overall accuracy was very high in this learning context, a period of assumed sleep better preserved new word memory than a period of assumed wake, allowing users to arrive at the correct foreign translation more quickly. This sleep-associated benefit was observed for vocabulary, but not for phrases, supporting previous suggestions that word-forms may be the aspect of language knowledge that benefits most from offline consolidation. We discuss this finding in the context of theoretical models of language learning, and consider the scope for using naturally occurring datasets to drive both research and practice.

According to the Complementary Learning Systems model, sleep permits the integration of new lexical representations into long-term vocabulary knowledge (Davis & Gaskell, 2009). Although experimental studies have frequently shown improvements in the number of words recalled across periods of sleep (e.g., Dumay & Gaskell, 2007; Henderson et al., 2012), we found no differences in overall memory accuracy between wake- and sleep-associated periods in this naturally occurring dataset. This null effect is likely because performance is extremely high in these types of data, with little room for variability (see Hopman et al., 2018, for discussion of similar issues). While this high accuracy is challenging from an analytical perspective, the format of the app is clearly motivating for users and facilitates engagement with learning. Understanding memory consolidation in these settings is thus as important as understanding what consolidation looks like in the lab.

Despite high overall accuracy, we did observe sleep-associated benefits in completion time, a measure that reflects the difficulty of retrieving the correct translation and the number of attempts required to arrive at the correct answer. Although circadian influences (e.g., fatigue) may partially account for the differences observed, the additional analyses favoured a predominant role for memory processes in driving sleep versus wake differences. This interpretation is also supported by experimental studies that have measured retrieval time for new word knowledge. For example, James, Gaskell, and Henderson (2020) found that children were quicker to name pictures after sleep, but not wake, and that sleep-associated improvements held across the following day. These converging lines of evidence support a role for sleep in accessing new words in memory, which can be observed across learning contexts.

Our primary hypothesis related to consolidation of new vocabulary, given that findings of sleep-associated benefits for word-form knowledge have been well-replicated in the lab. However, we also asked whether there were sleep-associated benefits for constructing phrases. Some studies have examined the role of sleep in extracting grammatical regularities (e.g., Mirković & Gaskell, 2016; Nieuwenhuis et al., 2013), but few have considered learners' ability to construct a phrase for meaning. In the *Memrise* phrase task, users were required to select the relevant words and order them to construct the correct phrase. We did not find evidence of a sleep-associated benefit for this more syntactic aspect of language knowledge. While challenging to conclude from in the present analyses (i.e., it does not constitute evidence in favour of the null hypothesis), our overall pattern of findings is consistent with evidence that sleep is most beneficial for consolidating new word-forms (e.g., Davis & Gaskell, 2009; James et al., 2020).

The present findings contribute to the vocabulary learning literature in three ways. First, we have demonstrated behavioural benefits of sleep-associated consolidation "in the wild", when users have naturally engaged in learning tasks at times to suit their schedule. This measure of naturalistic learning differs in that learners are genuinely motivated to learn the new information, and do so in the context of their everyday activities. As such, we demonstrate that behavioural benefits of sleep-associated consolidation are not restricted to tightly controlled experimental studies, and that they may be relevant to real-world learning goals.

Second, a key criticism of experimental studies is that they tend to focus on the first night of sleep-associated consolidation after learning new material. Whilst demonstrating clear changes in memory on this timescale, models of systems consolidation were founded on neuropsychological case studies demonstrating gradients of memory change over months and years (McClelland, McNaughton, & O'Reilly, 1995). Our analyses examined sleep-associated benefits across varied time-points in learning, suggesting that changes—although small—may be observable across prolonged periods of consolidation. We took an inclusive approach to analysing periods of sleep and wake regardless of prior experience with the item, but naturally occurring datasets will provide unprecedented opportunities for examining sleep-associated benefits across prolonged periods of consolidation in the future. Third, and relatedly, we showed that these benefits are observable despite mass testing and retrieval, otherwise proposed to

facilitate consolidation processes (Antony et al., 2017) and generally avoided in experimental studies of consolidation. Observing sleep-associated benefits in the context of repeated learning and testing is crucial for bridging gaps between studies of memory mechanisms and educational settings, as learners frequently receive feedback when the ultimate goal is to acquire new knowledge.

Challenges and Future Directions

We have demonstrated the potential for using naturally occurring datasets to examine theories of learning and consolidation. Like others (e.g., Hopman et al., 2018), we note the challenge of applying analyses to high accuracy data in app-based learning contexts, but we have also shown that behavioural differences can be observed in speed of access. These analyses make a key step towards practical applications of language consolidation research by demonstrating that tight experimental controls and highly salient settings are not critical to observing sleep-associated benefits. It is important to note, however, that this kind of dataset cannot provide information on sleep itself. In conducting these analyses, we assume that—in line with the broader population—the *majority* of users sleep during the night, and that atypical sleep patterns are a source of noise in the data. In light of such limitations, analysing naturally occurring data is best considered as *complementary* to lab-based research at present, rather than placed to determine the mechanisms of sleep-associated memory consolidation itself. The scope for integrating information across smartphone apps (e.g., activity trackers) presents an exciting direction for remediating these issues in the future.

Another key limitation is that learning is measured within the context of the app: it is unclear how speed of access benefits might translate to real-world language *use*. This uncertainty stimulates new—related—questions that could be explored in the lab. For example, changes in completion time could be driven by the precision of the memory—the more imprecise the representation is, the more repeated attempts will be required to reach the correct answer. Alternatively, completion time could reflect how accessible the relevant representation is to the user in the first place. These different underlying mechanisms could have different consequences for communicating in the real world, but have yet to be examined experimentally.

We believe the present analyses show promise for using naturally occurring datasets to complement experimental studies of vocabulary learning and consolidation (Goldstone & Lupyan, 2016). While the findings here will benefit from replication in another dataset, they also lay the foundations for future studies that could advance both theoretical and applied knowledge. Beyond the suggestions above, the data are well-suited to address questions of linguistic diversity and prior knowledge in learning. Moreover, there is scope for more closely examining whether there is a benefit for learning closer to bedtime, or whether sleep may be able to compensate for users that do not engage with the app as frequently. Our findings thus contribute a crucial first step towards capitalising upon naturally occurring datasets in this field, with potential for bridging the gap between experimental studies of language learning and language learning in practice.

Acknowledgments

This research was supported by an ESRC Postdoctoral Fellowship (ES/T007524/1) awarded to Emma James. We would like to thank Gareth Gaskell, Lisa Henderson, Scott Cairney, and Asifa Majid for helpful discussions and feedback throughout the project.

References

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*, 573-576.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1*.

Christensen, R. H. B. (2015). ordinal-Regression Models for Ordinal Data. R package version 2015.6-28.

Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21*, 803-820.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364*, 3773-3800.

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews, 13*, 309-321.

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., . . . Parsonage, H. (2019). Package 'data.table'.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18*, 35-39.

Fischer, S., & Born, J. (2009). Anticipated reward enhances offline learning during sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1586.

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory, 13*, 259-262.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in cognitive science, 8*, 548-568.

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, M. G. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science, 15*, 674-687.

Hopman, E., Thompson, B., Austerweil, J., & Lupyan, G. (2018). *Predictors of L2 word learning accuracy: A big data investigation.* Paper presented at the the 40th Annual

Conference of the Cognitive Science Society (CogSci 2018).

James, E., Gaskell, M. G., & Henderson, L. (2020). Sleep-dependent consolidation in children with comprehension and vocabulary weaknesses: It'll be alright on the night? *Journal of Child Psychology and Psychiatry*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*.

Lindsay, S., & Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning, 60*, 45-63.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419.

Mirković, J., & Gaskell, M. G. (2016). Does Sleep Improve Your Grammar? Preferential Consolidation of Arbitrary Components of New Linguistic Knowledge. *PloS One, 11*, e0152489.

Nieuwenhuis, I. L., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep promotes the extraction of grammatical rules. *PloS One, 8*, e65046.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods, 49*, 1630-1638.

Stafford, T., & Haasnoot, E. (2017). Testing sleep consolidation in skill learning: A field study using an online game. *Topics in cognitive science, 9*, 485-496.

Studte, S., Bridger, E., & Mecklinger, A. (2017). Sleep spindles during a nap correlate with post sleep memory performance for highly rewarded word-pairs. *Brain and Language, 167*, 28-35.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*: Springer.

Wilhelm, I., Diekelmann, S., Molzow, I., Ayoub, A., Mölle, M., & Born, J. (2011). Sleep selectively enhances memory expected to be of future relevance. *Journal of Neuroscience, 31*, 1563-1569.