Supplemental Online Material

Supplementary Methods

Participants (supplementary)

Recognition and generalization data for one wake participant were lost due to equipment failure. All were right-handed monolingual native-English-speakers. On the day of the experiment, participants were asked to wake before 8 am and refrain from napping or consuming any caffeinated products. Participants were pseudo-randomly allocated to sleep and wake groups, and received £10 for participation. The desired number of participants was determined (20 per group) with reference to previous studies that had used the speech error paradigm and that had shown correlations with sleep measures. Two no-shows left us with 19 participants per group. The research was approved by the Department of Psychology Ethics Committee at the University of York.

Design and Materials (supplementary)

The Test of Word Reading Efficiency (TOWRE) was administered to provide an indicator of participants' reading ability (Torgesen, Wagner, & Rashotte, 1999). Upon completion of the experiment, participants were given a short post-test questionnaire on their sleep habits over the previous 4 days in order to determine whether any participants had unusual sleep habits (none had). A subset of participants (N = 24) completed the Stanford Sleepiness Scale (Hoddes, Dement, & Zarcone, 1972) at four points during the experiment (before and after both training and testing) to provide a measure of any major changes in sleepiness and motivation that may have affected performance. These participants were also given a post-test to determine whether they had noticed any rules in the items they were presented.

Error Coding

Production errors were coded as in previous studies of this kind (e.g., Warker, 2013) using two error types (same-position and different-position) applied to the restricted (/f/, /s/) and unrestricted (/k/, /g/, /m/, /n/) consonants. An error in which a slipping consonant retained the same syllable position in a given sequence (e.g., mistakenly saying *sang sam* instead of *sang gam*) was described as a same-position error whereas an error that involved a change of syllable position (e.g., mistakenly saying *sang gas* instead of *sang gam*) was classified as a different-position error. Most speech errors tend to preserve the syllable position of the slipping consonant, although of course individuals vary in their adherence to this constraint. For unrestricted consonants, the percentage of all errors (same-position + different-position) that were same-position errors establishes a baseline for this tendency in the absence of any new constraints. For example, a typical participant in our experiment (see Table 1 of the main article) might make 45 speech errors involving unrestricted consonants in the course of the training session. If 7 of those errors involved a change of position, with the remaining 38 being same-position errors, then the same-position percentage for that participant would be 38/45 = 84%. For restricted consonants, a value for this percentage that is higher than the baseline level for unrestricted consonants would indicate that the participant's speech errors were adhering to the within-experiment constraints (given that when errors occurred, the uttered syllable always had the same vowel as the intended syllable). For example, if the same participant in training made 85% same-position errors on the restricted consonants during training then their phonotactic learning score at that point in the experiment would be 85%-84% = 1%. Coding was carried out by one of the authors, with cross-checking of a subset of error recordings by two other coders showing high concordance rates. When the primary coder found no error, the agreement rate was

97%, whereas when the primary coder found an error the agreement rate was between 71% and 78%, similar to other studies using this methodology.

Sleep Recording

Sleep participants were wired up using 9 EEG montage (F3-A2, F4-A1, C3-A2, C4-A1, O1-A2, O2-A1) using Ag-AgCl cup electrodes, which were applied according to the international 10-20 system. Frontal and central electrodes were selected in order to observe slow-wave and spindle activity, and delta power analyses were carried out across all four electrodes. Occipital electrodes were used to observe alpha activity for visual scoring of sleep, but were not further analysed. Two electro-oculogram electrodes and two electromyogram electrodes were also applied. Impedance levels, sampling rates and filter settings were set according to the American Academy of Sleep Science Manual (Iber, 2007).

Sleep Scoring

Sleep data were recorded digitally at a sample rate of 200 Hz using an Embla N7000 system with RemLogic 3.0. Sleep data for each participant were manually scored in RemLogic by two independent coders. Data were scored in 30-second epochs using standard criteria (Iber, 2007), with close inter-scorer agreement. Sleep spindle data were analyzed, but did not show a significant correlation with the key behavioral results.

Supplementary Results

Potential confounds: TOWRE

Performance on the TOWRE did not correlate with any sleep stage or measure of power, but given modest non-significant correlations between this variable and measures of slow-wave sleep (SWS minutes, $r = .30$; delta power, $r = .37$; SWA, $r = .37$) it is worth noting that the

correlations between the slow-wave sleep measures and the change in speech error patterns all remained significant when TOWRE performance was partialed out.

Potential confounds: Sleepiness

For the subgroup who provided sleepiness scale data, the scores were analyzed using a three-way ANOVA with a between-participants variable, Group (sleep vs. wake), and two within-participants variables: Session (training vs. test) and Order (before vs. after the tests). This analysis revealed a marginal main effect of Session [$F(1, 19) = 3.9$, $p = .064$, $\eta_p^2 = .17$] and two significant interactions. All other effects and interactions did not approach significance. The interaction between Order and Session [$F(1, 19) = 4.5$, $p = .048$, $\eta_p^2 = .19$] was not of theoretical value, but the Session x Group interaction [$F(1, 19) = 15.3$, $p < .001$, $\eta_p^2 = .45$] was more relevant to the key behavioral effects. Whereas wake participants showed a non-significant increase in their rated sleepiness (from 2.8 to 3.1) between the two sessions, the sleep group showed a significant reduction in their ratings (from 3.2 to 2.3). Given that the second session was after the sleep group had rested, this change was not surprising, but it could perhaps be argued that the change in the type of speech errors for this group was a consequence not of memory consolidation during sleep but somehow of the greater alertness that this group had after sleep. We addressed this possibility in several ways. First, we calculated the change in sleepiness across sessions for both wake and sleep participants and correlated this with the two key behavioral effects that we found. Neither correlation was significant (phonotactic measure: $r = .343$, $p = .13$; generalization: $r = .19$, $p = .41$). Although these non-significant effects are partly reassuring, it is hard to rule out a potential confound on the basis of failure to reject the null hypothesis. Consequently a second form of analysis examined whether, for the sleep group, the correlations between the key sleep variables and performance changes held when changes in

sleepiness were partialed out. The significant correlations between slow-wave sleep measures and the phonotactic effect in fact got numerically stronger rather than weaker when controlling for changes in sleepiness (SWS Duration: r = .69, p = .029; Delta Power: r = .66, p = .038). A similar result was obtained using logistic regression. We examined the model containing the 2-way interaction between Session and Restriction, the 3-way interaction between Session, Restriction and SWS Duration, the 3-way interaction between Session, Restriction and Sleepiness Change, and the 4-way interaction between all these variables, along with an intercept. If sleepiness was underlying the change in performance then we would expect interactions involving this variable to be significant and not interactions without this term. In fact only the Session x Restriction x SWS Duration interaction showed an independent effect on speech errors (β=0.07, SE=.04, z=2.03, p=.04). Equivalent analyses replacing SWS Duration with Delta Power produced the same pattern of results (Session x Restriction x Delta Power: β=0.01, SE=.005, z=2.08, p=.04). In sum, there was good evidence that the change in performance for the sleep group was associated with structural properties of the nap rather than changes in sleepiness.

Explicit knowledge

The same subgroup were asked to write down any rules that they could determine about the items in the experiment. Some participants noticed that the vowels alternated between *a* and *i* and some noticed that the items began and ended with a consonant. One participant incorrectly thought that the *i* sequences had more *f* and *s* consonants. However, most participants did not volunteer any observations, and none of them noticed any association between consonants and vowels. Hence, as in previous studies, the newly extracted knowledge was implicit in that it was not describable.

References

Hoddes, E., Dement, W., & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiology, 9*(150), 431-436.

Iber, C. (2007). *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*: American Academy of Sleep Medicine.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Test of word reading efficiency: Austin, TX: Pro-Ed.

Warker, J. A. (2013). Investigating the retention and time course of phonotactic constraint learning from production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 96.