Effect of similarity and training experiences on new vocabulary learning

Megan Waller (meganwal@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Daniel Yurovsky (dyurovsky@gmail.com)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Nazbanou Nozari (bnozari@iu.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E 10th St, Bloomington, IN 47405

Abstract

In two experiments (N = 179), we studied the effect of contextual similarity and training mode on new vocabulary learning. Adult participants were trained on blocks of items that were semantically similar, phonologically similar, or unrelated to one another. Each participant was trained through passive exposure, active comprehension, or active production of the new vocabulary. Exp 1 trained items in clusters of 9, whereas Exp 2 trained the same number of items in clusters of 3. Exp 2 also assessed delayed retention 48-72 hours after training. Results showed a robust and negative impact of semantic similarity and production mode on vocabulary learning. A detrimental effect of phonological similarity was only observed in the delayed test. These results suggest that adding the challenge of resolving similarity-induced competition and articulating the word-form negatively impacts the quick acquisition of new vocabulary.

Keywords: vocabulary learning; word production; contextual similarity; semantic interference; phonological interference

Introduction

Although learning a new language has many facets, learning the relationship between words and the meaning they specify, i.e., vocabulary learning, is a basic block in the process. Naturally, a central question in research on second language learning is "What is the most efficient way to teach/learn new words?". Past research has identified a few principles for enhancing learning. For example, learners benefit from spacing (interleaving training items) over massed practice (repeated studying of the same item; Kornell, 2006), and from being tested over repeated study (Roediger & Karpicke, 2006). Specifically in language, two issues have caused debates: learning mode and contextual similarity among the to-be-learned items. Learning mode refers to the training method and its interaction with learning goals. For example, if the goal is for the listener to be able to comprehend new words, will learning be better if the training method emphasizes comprehension or production? Contextual similarity refers to the relationship between items in a training set. Most pedagogical settings group new words into semantically related categories, e.g., "animals", "clothing items", "fruits", etc. But does semantic similarity facilitate or hinder learning? How about phonological similarity? Is it easier or harder to learn similar-sound words such as "cap", "map", "cat", "mat" together in one set?

Investigations of both learning mode and contextual similarity are highly motivated by past findings in cognitive research. For example, the principle of *desirable difficulty*, uncovered by research in memory, posits that making learning more challenging should benefit long-term retention of information, i.e., learning. In keeping with this prediction, Hopman and MacDonald (2018) reported better learning of a new language when training demanded participants to produce words (production mode) vs. when training only required them to listen to and comprehend words (comprehension mode). Similarly, several studies have shown that participants are slower and more error prone in naming pictures in the context of semantically or phonologically related items (e.g., Schnur et al., 2006; Breining, Nozari & Rapp 2016; Nozari et al., 2016). One explanation for this finding is that it results from incremental learning processes (Oppenheim, Dell & Schwartz, 2010; Oppenheim & Nozari, 2024). If incremental learning is indeed the underlying mechanism for contextual similarity interference, we can expect training in similar context to yield poorer results. This prediction was partially supported by Korochkina, Bürki, and Nickels (2021) who reported poorer learning of a novel language in a semantically related context.

This brief literature review demonstrates the critical importance of investigating the roles of learning mode and contextual similarity in new vocabulary learning. However, there is currently both controversy and unanswered questions. For example, Leach and Samuel (2007) trained participants on novel words with a word-picture matching task where some participants produced the target word aloud afterwards. In this case, the difficulty added by producing the word hindered accuracy in a perceptual categorization task. This study, however, evaluated learning on the perceptual knowledge of word forms rather than comprehension.

Similarly, studies that investigated the role of semantic similarity have yielded mixed results, ranging from facilitation (Hoshino, 2010) to null results (Nakata & Suzuki, 2019) to interference (Nozari et al., 2016). The role of phonological similarity has been investigated less rigorously. A negative influence of phonological overlap was found on

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

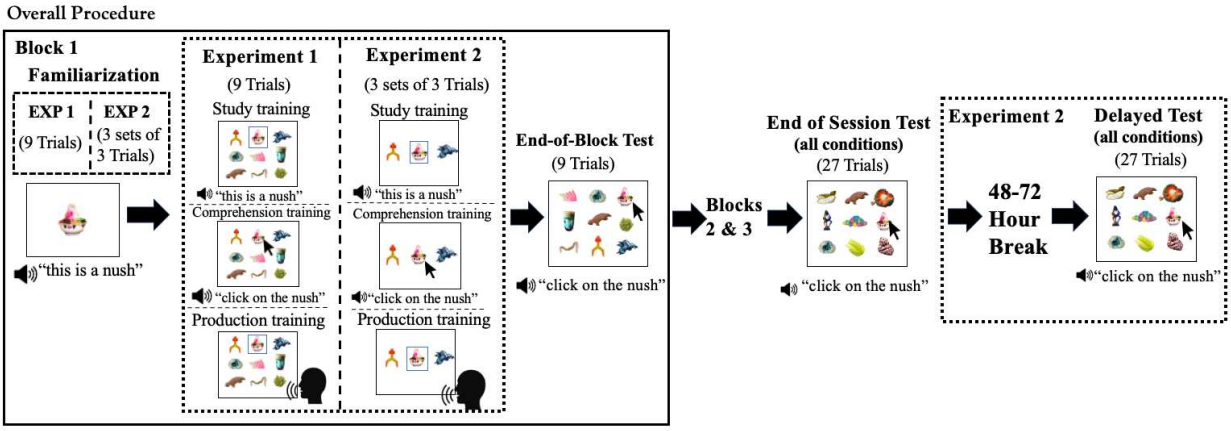

Figure 1. Overall Procedure of Experiments 1 and 2. Differences between Exps 1 and 2 are shown within dotted lines.

learning to write new words, but this effect could have a strong orthographic component (Breining, Nozari & Rapp, 2019). Moreover, while the effects of the two factors have been studied separately, it remains unclear whether they modulate each other. Finally, the more rigorous designs, e.g., Korochkina et al. (2021) did not measure delayed retention.

This study addresses these issues. In two experiments, we trained a total of 179 participants on 27 new vocabulary items from an artificial language. Training mode and contextual similarity were each manipulated with three levels. Each participant was assigned either to a study (passive listening), a comprehension (listening and selecting), or a production (speaking) training mode. All participants completed blocks of semantically related, phonologically related, or unrelated items. Learning was assessed after each block, as well as at the end of the session. Exp 1 and 2 differed in the difficulty of initial training: Exp 1 trained all nine items within the block simultaneously, whereas Exp 2 broke them in clusters of three. Finally, Exp 2 added a delayed test (48 to 72 hours after the end of training) to assess longer-term retention.

Experiment 1

Method

Participants For sample size estimation, given no prior study has examined this interaction, we set the effect size to medium (Cohen's d = 0.5) and conducted a power analysis using PANGEA (Version 0.2; Westfall, 2016). With 30 participants per training mode (N = 90) and 9 items per similarity condition, this simulation yielded 80.8% power to detect the critical interaction.

Anticipating possible attritions, 98 participants were recruited online through Prolific. Participants were all native English speakers from the United States or Canada. Eight were removed for technical issues. The remaining 90 participants (ages 20-40 years, mean age 30.74 years; 66.7% men, 31.1% women, and 2.2% non-binary) were randomly assigned to one of the three learning modes.

Materials Materials consisted of six sets of nine images and three sets of nine words. Images of unfamiliar objects were selected from Google Images and Novel Object & Unusual Name Database (Horst & Hout, 2014). Three of the sets each formed a semantic category: birds, flowers, and fruits. For the unrelated sets, one was used for practice trials, and the other two were assigned to the unrelated or phonological similarity blocks. For all image sets, visual differences in color, orientation, and shape were balanced as much as possible.

We created 27 novel words and divided them into three sets, balanced for syllables and phonemes. To quantify phonological similarity, we used position-independent phonological overlap, defined as the total number of phonemes shared by two strings, regardless of position, divided by the total number of phonemes in the two strings (Goldrick et al., 2010). We calculated overlap between all pairs, then averaged across pairs in a set. Phonological similarity scores were 0.088 and 0.085 for the two unrelated sets, and 0.383 for the phonologically related set. This difference is comparable to studies that found interference effects in language production (Breining et al., 2016). Audio recordings were generated with an artificial voice program, Descript (https://www.descript.com/).

Label-image mappings were pseudo-randomly generated for each participant. One of the three semantic image sets was randomly paired with one of the unrelated label sets for the semantic block, one of the unrelated image sets was randomly paired with the phonological label set for the phonological block, and then the remaining two unrelated sets created the unrelated block. Within each block, the images were randomly mapped to labels for each participant.

Procedure The task was conducted online using JavaScript code with JsPsych plugins. Participants were asked to learn novel labels for 27 pictures. Two sets of factors were manipulated, training mode and contextual similarity, each with three levels. Mode consisted of study, comprehension, and production, manipulated between subjects. Similarity consisted of semantic, phonological, and unrelated blocks, manipulated within subjects.

Figure 1 shows the overall structure of a session. After consenting and orientation, participants completed three practice trials similar to the experimental blocks (see below).

Participants were then presented with three blocks (semantic, phonological, and unrelated in randomized order) which each consisted of familiarization phase, a training phase, and an End-of-Block test. The familiarization and test phases were identical across modes. During familiarization, participants were shown each image in a random order, heard their labels (e.g., "This is a nush"), and were asked to repeat the word aloud before pressing a "continue" button to proceed. Next, they moved on to training where participants learned the labels differently depending on their mode. All trials began with a 3x3 grid of all images in a block, where the position of images changed on each trial. In the study mode, participants listened passively. After 2000 ms a blue border appeared around the target image and the correct label played aloud (e.g. "this is a nush"). 2000 ms later, the border turned green, and the participant clicked on that image to proceed. In the comprehension mode, after 2000 ms, participants were instructed to select the corresponding image (e.g. "click on the nush"), then a blue border confirmed their selection. A green border then outlined correct image as feedback before moving on. In the production mode, after 2000 ms, a blue border appeared around the target image, and participants were asked to say the label or their best guess, then click on the image to hear the correct label aloud and move on. For all modes, there was no response deadline, but participants were reminded to respond after 5000 ms.

Immediately after training, participants completed an End-of-Block test: a word-to-picture matching test to measure comprehension learning. Participants saw a 3x3 grid of all nine pictures from this block, after 2000 ms heard one of the trained labels, and then clicked on the appropriate image. A blue border confirmed their selection, then proceeded to the next trial without feedback. There was again no response deadline, but participants were reminded after 5000 ms.

Finally, after completing all three blocks, they completed the End-of-Session test. The procedure was identical to the End-of-Block comprehension test, except the grid contained images from *all* blocks (three images from each block per trial), with one trial for each of the 27 labels. Each of the 27 images appeared as an option in nine trials in a random order.

Data Processing Accuracy of comprehension trials and reaction times (RTs) were automatically recorded, and production trials were transcribed by hand and coded for whether the participant gave an accurate label (Strict) or missed only one phoneme (Lenient). Providing no response was coded as an error. RTs were only analyzed for correct trials, and RTs for any trials where the participant responded three standard deviations above or below their mean were removed. RTs were log-transformed for the analysis.

Statistical Analysis Unless stated otherwise, all analyses were carried out using (general) linear mixed effect models with *lme4* package (Bates et al., 2015) in R (version 4.2.1, R Core Team, 2022). For training mode, we were interested both in the effect of the active modes of training (comprehension and production) against the passive study mode, and in directly comparing the active modes. So, we ran two sets of models for each time point. In the full model the study mode is the baseline to which comprehension and production are compared. The second (direct comparison) model only included data from the active training conditions, and comprehension is coded as the baseline. For contextual similarity, the two dimensions of similarity (semantic and phonological) are not comparable, so in both models, each similarity condition is compared against the baseline unrelated condition. We initially aimed for including the maximal random effect structure (Barr et al. 2013), but for consistency included the random intercept for subject and item, with which all models converged. The exact same structure was used for accuracy and RT models, except a logistic version of the model was used for the former with a binary accuracy measure, with incorrect as 0 and correct as 1.

Results

Within-Block-Training Figure 2a shows the accuracy during the Within-Block training, using lenient coding for production.[1] Study mode were passively listening and could not make errors. A Mann-Whitney U test comparing participants' mean accuracy by mode showed participants were less accurate during production than comprehension (M_{prod} = 23%, M_{comp} = 53%, W = 817, $p < 0.001$).

As for effects of contextual similarity, we conducted Wilcoxon Signed Ranks test comparing mean accuracy across participants for each similarity condition and the unrelated condition. Participants were less accurate in the semantic block compared to unrelated block (M_{Sem} = 28%, M_{Unrel} = 41%, V = 1029, $p < 0.001$), but there was no difference between phonological and unrelated block accuracy (M_{Phon} = 44% M_{Unrel} = 41%, V = 409.5, p = 0.44).

End-of-Block Tests Figure 2b shows the accuracy and RTs on the End-of-Block tests. All models had mode, similarity and their interaction as fixed effects and random intercept of subjects and items as random effects. For mode, accuracy in the production was marginally lower than study (z = -1.87, p = 0.061). For similarity, there was poorer accuracy in the semantic compared to unrelated similarity condition for both comprehension (z = - 3.49, $p < 0.001$) and production modes (z = -2.481, p = 0.013) compared to study. In the direct comparison model, accuracy was significantly lower in the semantically related condition (z = -4.454, $p < .001$). No other effects were significant in the accuracy or the RT models, removing concerns regarding a speed-accuracy tradeoff.

End-of-Session Test Figure 2c shows accuracy and RTs on the End-of-Session test. This time, production training

[1] There were no differences in results between Strict and Lenient coding for any analyses, therefore only results for Lenient coding are reported.

participants showed significantly poorer accuracy than study (z = -2.436, p = 0.015; full model), and comprehension (z = -2.356, p = 0.019; direct comparison model). There also was lower accuracy in semantically related compared to the unrelated condition in the active mode comparison model (z = -2.498, p = 0.013). None of the other effects on accuracy were significant. In the full RT model, the interactions between comprehension (t = 2.072, p = 0.038) and production (t = 2.404, p = 0.016) modes and semantic similarity were both significant, suggesting slower responses in the semantic compared to unrelated similarity condition for both active modes compared to study. The RT model directly comparing active modes revealed longer RTs in the semantic compared to the unrelated condition (t = 3.57, p < 0.001), and an interaction between production and phonological similarity, where RTs were longer for phonological than unrelated condition in production mode (t = 2.206, p = 0.028).

Discussion

As expected, accuracy during training was lower in the production than the comprehension mode, and participants were less accurate while learning labels in the semantically, but not phonologically, related condition. The End-of-Block test revealed less accurate performance on semantically related, compared to unrelated, blocks when embedded in production and comprehension modes, compared to study. When production was compared directly against comprehension, the model showed significantly lower accuracy in the semantic condition. For mode, accuracy in the production mode was marginally lower than study, but not significantly different from comprehension.

The most critical test of the experiment, however, is the combined test, which measures learning of all trained items in a mixed context. Here, there was clear evidence that learning in the production mode was less accurate than both other modes. Semantic similarity also had a detrimental effect in these two active modes of learning: overall accuracy was lower in the semantic compared to the unrelated condition in the model that included data from these two modes. Also, RTs were significantly slower in the semantic vs. unrelated condition for both production and comprehension modes vs. the study mode. The effect of phonological similarity on learning was much less robust. We only observed a disadvantage in RTs for learning in production compared to comprehension mode. To summarize, these results suggest a negative impact of the production mode on learning vocabulary in perception. They further show that semantic similarity among the items in the training set can be detrimental to learning in active learning modes.

Exp 2 followed two goals. First, it was designed to provide a conceptual replication of Exp 1; we tested whether the disadvantages observed for the production mode and semantic similarity were robust enough against the details of the training scheme. In Exp 1, all nine items within a block were presented simultaneously, leading to relatively low performance during training, especially in the production mode (23%). However, many modern learning apps, e.g.,

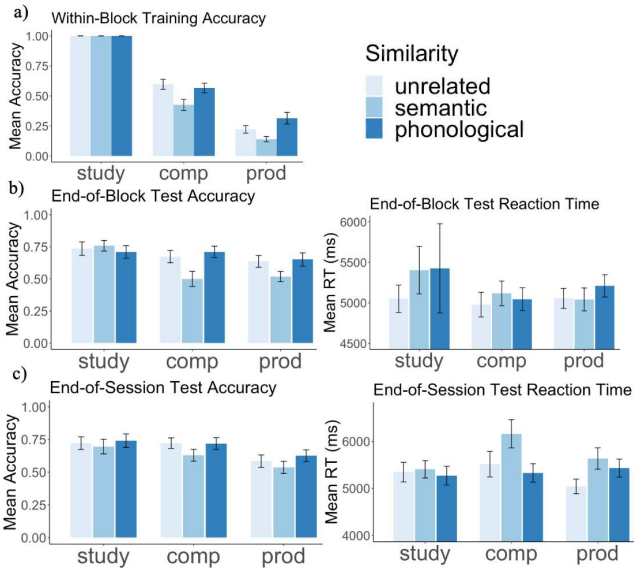

Figure 2. Accuracy (left) and RTs (right) for a) training, b) End-of-Block tests and c) End-of-Session test in Exp 1. Error bars show SEs.

Babble, train items in smaller clusters of 3 or 4 items. In Exp 2, a new group of participants were still assigned to three modes and each completed training in three similarity contexts. However, within each block, items were trained in three triplets, which we expect to lead to better within-block performance. Does this change modify the negative impact of production and semantic similarity on learning? The second goal of Exp 2 was to test the longer-term effects of mode and similarity on learning, by testing again 2-3 days after training. This delayed assessment provides a further test of the robustness of the reported effects on learning.

Experiment 2

Method

Participants Sample size estimation was the same as Exp 1. Anticipating possible attritions, 101 participants were recruited online through Prolific. Participants were all native English speakers from the United States or Canada. Twelve participants did not complete the experiment. The remaining 89 participants (ages 18-40 years, mean age 31 years, 50.6% men, 47.2% women, 1.1% nonbinary) were randomly assigned to one of three learning modes: 30 to study and comprehension, 29 to production.

Materials The same word and image sets as Exp 1 were used, except one semantic image set (birds) was removed to accommodate changes in the word-image mapping procedure. Unlike in Exp 1 where mappings were randomly determined for each participant, to ensure the mappings were the same on both days of the experiment, participants were randomly assigned to one of four mapping lists, counterbalanced across training modes. These lists balanced

the matching of each word set with each image set, with the pairings of words and images randomly determined.

Procedure This Experiment was also conducted online using JavaScript code with JsPsych plugins. Figure 1 shows the differences between Experiment 1 and 2. Exp 2 was different from Exp 1 in two ways. First, within each block, rather than complete familiarization and training with all nine words at once, the nine items were divided into three triplets, and participants completed familiarization and training with each triplet separately, the order randomized for each participant. Which three images appeared together were pre-determined to control for visual similarity across conditions and sets. After familiarization and training were completed, the end of block test completed an End-of-Block test identical to Exp 1 containing all 9 items. After all three blocks were completed, they completed an End-of-Session comprehension test that mixed together items from all blocks, identical to Exp 1.

The second difference between Exp 1 and Exp 2 was the addition of a delayed test 48-72 hours after completing the first session. Participants completed another comprehension test identical to the End-of-Session test on the first session, with the 27 trials presented in randomized order.

Data Processing and Statistical Analysis Data processing and statistical analysis procedures were identical to Exp 1.

Results

Within-Block-Training. Figure 3a shows the accuracy during the Within-Block training phase. We followed the same procedures as Exp 1, and the results were similar: participants were significantly less accurate during production than comprehension training (M_{prod} = 57%, M_{comp} = 89%, W = 813.5, p < 0.001) and performed worse in the semantic block compared to unrelated block (M_{Sem} = 69%, M_{Unrel} = 75%, V = 577.5, p = 0.023). There was no significant difference between phonological and unrelated block accuracy (M_{Phon} = 77% M_{Unrel} = 75%, V = 241, p = 0.226).

End of Block Test Figure 3b shows the accuracy and RTs on End-of-Block tests. Both the accuracy and RT model structures were identical to the models used for analyzing Exp 1. In the full model, participants showed poorer accuracy for identifying items from the semantic similarity block across all modes (z = -3.217, p = 0.001). In the model comparing production and comprehension modes directly, there was also a significant main effect of semantic similarity (z = -4.135, p < 0.001), and marginally poorer accuracy in production compared to comprehension (z = -1.919, p = 0.055). No other effects on accuracy or RTs were significant.

End-of-Session Test Figure 3c shows the accuracy and RTs on the End-of-Session test. Participants again performed worse on items with semantic similarity across all modes (z = -2.35, p = 0.019), and in the direct comparison model (z = -3.329, p = 0.001). Participants trained in production mode performed significantly worse across all similarity conditions compared to study (z = -2.679, p = 0.007), and when

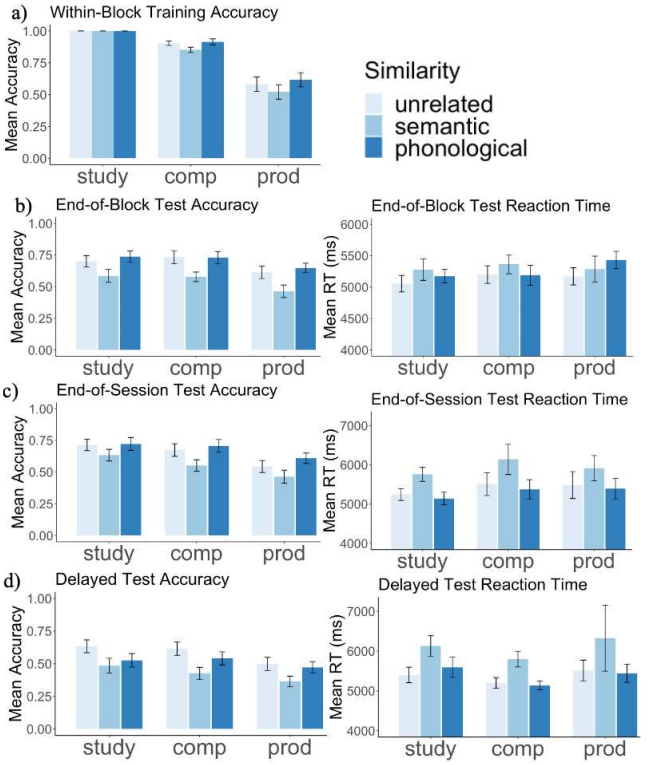

Figure 3. Accuracy (left) and RTs (right) for a) training, b) End-of-Block test c) End-of-Session Test and d) Delayed Test in Exp 2. Error bars show SEs.

compared directly to comprehension (z = -1.982, p = 0.047). Each of the RT models showed a robust disadvantage for semantic similarity (t = 3.756, p = < .001; t = 3.88, p = < .001, for the full and direct comparison models respectively). No other effects were significant.

Delayed Test Figure 3d shows accuracy and RTs for the Delayed test conducted 48 to 72 hours after training. In the full model, participants showed poorer accuracy for both semantic (z = -3.95, p < 0.001) and phonological (z = -2.221, p = 0.026) overlap compared to the unrelated context. There was again a negative main effect for the production compared to study mode (z = -2.068, p = 0.039). When comparing active learning modes directly, there was again a significant disadvantage of semantic similarity (z = -4.921, p = <.001), as well as marginally poorer accuracy in production compared to comprehension (z = -1.982, p = 0.069). The RT models showed a robust disadvantage for semantic similarity (t = 4.281, p = < .001; t = 4.253, p = < .001, full model and comparison model respectively). Additionally, the RT models showed a marginally significant interaction between semantic similarity and production mode in the full model (t = -1.944, p = 0.052) and a significant interaction between the two in the direct comparison model (t = -2.177, p = 0.03).

Discussion

Although, as expected, within-blocking learning was easier in Exp 2, the results were largely similar to Exp 1 and, in

some ways, cleaner. For example, the detrimental effect of semantic similarity on the End-of-Block test in Exp 1 was observed in the interactions between production/ comprehension and similarity, but it showed up as a main effect in Exp 2, suggesting a general effect that held across all modes of learning. Importantly, the results of the End-of-Session testing were replicated: there was a robust detrimental effect of production compared to both study and comprehension modes. Moreover, Exp 1 only found a negative effect of semantic similarity on accuracy in the model that only contained production and comprehension data, there was again a main effect of accuracy across the board in Exp 2, which was also reflected in slowed RTs in this condition compared to the unrelated condition. Contrary to Exp 1, however, there was no effect of phonological similarity on learning in the End-of-Session test.

Exp 2 also tested the retention of information 48-72 hours after training. This test again confirmed the detrimental effect of semantic similarity on learning across the board. Moreover, learning was significantly worse in the production mode compared to both study and comprehension. The delayed test also showed a detrimental effect of phonological similarity, but only in the full model, and not the model that only included the production and comprehension data sets.

Combined Analysis of Experiments 1 and 2

Given the similarities between the designs and the pattern of results in Exps 1 and 2, we conducted a combined analysis of the two datasets with greater statistical power. Since Exp 1 did not have a delayed test, this analysis was conducted on the End-of-Session test. These models had the same structure as before, but added Experiment as a fixed effect. There was no main effect of Experiment in any model, showing comparable levels of accuracy and RTs across the two, further supporting pooling the data. In the full model, semantic similarity (z = -2.277, p = 0.023) led to worse accuracy than the unrelated context, and production mode led to worse accuracy compared to study (z = -3.626, p < 0.001). The model directly comparing active modes mirrored these results with negative impact of both semantic similarity (z = -4.092, p < .001) and production mode (z = -3.063, p = 0.002). The corresponding RT models also found a significant detrimental effect of semantic similarity (z = 3.285, p = 0.001; z = 5.232, p < .001, full and direct comparison model respectively). Other effects did not reach significance.

General Discussion

In two experiments, we trained a total of 179 participants on novel vocabulary, and assessed the effects of contextual similarity and training mode on the acquisition of the new words. Despite differences in difficulty during the initial learning phase, the results of the two experiments were largely consistent, albeit with some minor differences. While there was evidence that semantic similarity and production mode may have a detrimental effect on learning, these effects were sometimes observed across all three modes, and sometimes only in the more active, production and comprehension modes. Also, some, but not all tests suggested a detrimental effect of phonological similarity on learning. Importantly, these effects were preserved in a delayed test conducted 48-72 hours after the end of training.

The similarity of the designs and the general pattern of results across the two experiments allowed us to conduct a combined analysis, which, to our knowledge, uses the largest sample exploring the joint effects of contextual similarity and training mode on vocabulary learning.

The results showed a robust and negative influence of semantic similarity on learning across all modes, as well as a negative effect of the production mode compared to both study and comprehension modes. The combined model, however, did not show a robust influence of phonological similarity on learning.

The finding of a negative impact of semantic similarity on learning replicates the report of Korochkina et al. (2021) and extends that to delayed testing. Theoretically, the finding fits well with the incremental learning accounts of semantic interference (Oppenheim et al., 2010; Oppenheim & Nozari, 2024), and the previous reports on the longevity of semantic interference (Hepner & Nozari, 2020). In contrast to semantic similarity, the effect of phonological similarity was not robust, and was only significant on the delayed test. A previous study that manipulated phonological similarity found a detrimental effect on novel vocabulary learning, but that study elicited written responses, which adds orthographic similarity to the mix (Breining et al., 2019). In general, while there is now sufficient evidence to support the presence of phonological interference in production (e.g., Breining et al., 2016; Qu, Feng & Damian, 2021), the effect is more elusive than semantic interference, being more sensitive to strategies such as noticing common onsets (O'Seaghdha & Frazer, 2014). Such strategies can lead to short-term facilitation (Nozari et al., 2016), which may explain the late emergence of the effect in the delayed test.

The detrimental effect of the production mode on learning aligns with studies of perceptual learning (Leach & Samuel, 2007; Baese-Berk & Samuel, 2016) but in contrast to the study of Hopman and MacDonald (2018) who reported better performance on comprehension tests for people who had engaged the production system in learning. One prominent difference between our study and that of Hopman and MacDonald is the focus on individual words in ours vs. sentences in theirs. In fact, when Hopman and MacDonald (2018) tested the learning of individual items, there was no advantage for the production mode. It thus remains possible that production is most advantageous for learning syntax, whereas the extra difficulty that is often associated with production hurts the quick acquisition of new vocabulary items.

Finally, our results showed that the effects of interest show up early, i.e., during learning, and persist over time, at least for 72 hours. In fact, the phonological effect was most prominent at this late point, suggesting a potential role for consolidation. Studying retention at later points is a great avenue for future research.

Acknowledgements

We thank Nikhil Lakhani for his assistance with the JavaScript code. This work was supported by the James S. McDonnell Foundation Scholar Award in Understanding Human Cognition #220020506 to D.Y., Spencer Foundation grant #202000221 to N.N, and NSF-BCS- 206378 to NN.

References

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36.

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*(3), 255-278.

Bates, D., Maechler, M., & Bolker, B. (2015). Walker., S. Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*(1), 1-48.

Breining, B., Nozari, N., & Rapp, B. (2016). Does segmental overlap help or hurt? Evidence from blocked cyclic naming in spoken and written production. *Psychonomic bulletin & review*, *23*, 500-506.

Breining, B., Nozari, N., & Rapp, B. (2019). Learning in complex, multi-component cognitive systems: Different learning challenges within the same system. *Journal Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1093–1106.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113-134.

Hepner, C. R., & Nozari, N. (2020). The dual origin of lexical perseverations in aphasia: Residual activation and incremental learning. *Neuropsychologia*, *147*, 107603.

Hopman, E. W. M., & MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, *29*(6), 961–971.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior research methods*, *48*, 1393-1409.

Hoshino, Y. (2010). The Categorical Facilitation Effects on L2 Vocabulary Learning in a Classroom Setting. *RELC Journal*, *41*(3), 301–312.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(9), 1297-1317.

Korochkina, M., Bürki, A., & Nickels, L. (2021). Apples and oranges: How does learning context affect novel word learning? *Journal of Memory and Language*, *120*, 104246. https://doi.org/10.1016/j.jml.2021.104246

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive psychology*, *55*(4), 306-353.

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*(2), 287–311.

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience*, *31*(7), 886–903.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010) The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227-252.

Oppenheim, G. M., & Nozari, N. (2024). Similarity induced interference or facilitation in language production reflects representation, not selection. *Cognition.*

O'Séaghdha, P. G., & Frazer, A. K. (2014). The exception does not rule: attention constrains form preparation in word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 797.

Qu, Q., Feng, C., & Damian, M. F. (2021). Interference effects of phonological similarity in word production arise from competitive incremental learning. *Cognition*, *212*, 104738.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199-227.

Westfall, J. (2016). PANGEA: Power ANalysis for General Anova designs. https://jakewestfall.shinyapps.io/pangea/