The effects of type and token frequency on semantic extension

Introduction

Methods

Following Harmon & Kapatsinski (2017), two artificial languages were used: Dan and Nem (See Figure 1). In each language, the same four suffixes were used: *-sil*$_{PL}$, *-dan*$_{PL}$, *-nem*$_{DIM}$, and *-shoon*$_{DIM}$. Notably, in our language *-dan* and *-sil* overlap in meaning (they both occur in plural contexts), and *-nem* and *-shoon* also overlap in meaning (they both occur in diminutive contexts). Since all four suffixes are possible candidates for the diminutive plural meaning, we will examine how properties of the language affect which suffixes are extended to express the diminutive plural meaning.

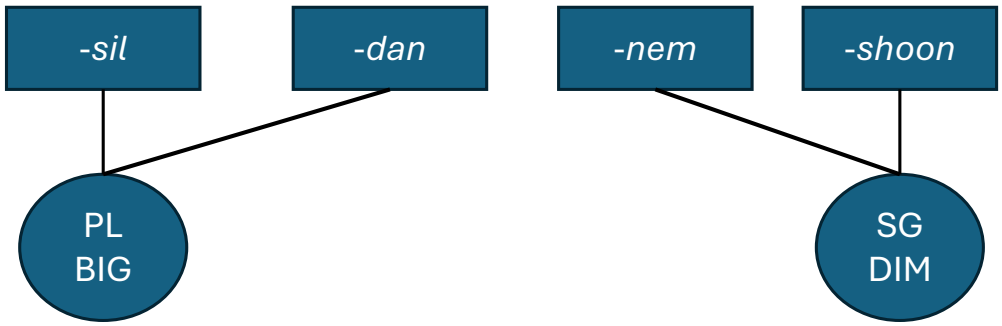

Figure 1: A description of the suffixes in our artificial languages. The thicker lines denote the more frequent form in each language: the plural *-dan*$_{PL}$ in the Dan language and the diminutive *-nem*$_{DIM}$ in the Nem language.

During the exposure phase, each suffix was was paired with an image. The suffixes *-sil*$_{PL}$ and *-dan*$_{PL}$ were always paired with a picture of multiple large pictures. On the other hand, the suffixes *-nem*$_{DIM}$ and *-shoon*$_{DIM}$ were always paired with a picture of a single small creature. The design of the stimuli results in participants being able to learn that *-sil* and *-dan* are either simply plural or

Table 1: Description of each of our conditions. Note that in Condition 1, there are an equal number of tokens between the frequent and infrequent items, however there are a greater number of types in the frequent items. In Condition 2, the opposite is true: the frequent items occur more, but in the same number of types as the infrequent items. Finally, in Condition 3, the frequent items occur both a greater number of times and in a greater number of different contexts.

	Frequent.Token	Frequent.Type	Infrequent.Token	Infrequent.Type
Type Frequency	12	12	12	3
Token Frequency	12	3	3	3
Type-Token Frequency	12	12	3	3

simply non-diminutive. Similarly, *-nem* and *-shoon* can be learned as either simply singular or simply diminutive.

Our Experiment comprised of three different conditions (see Table 1), one in which the type frequency of the frequent language's suffix was manipulated (Type Frequency), one in which the token frequency was manipulated (Token Frequency), and one in which both were manipulated (Type-Token Frequency). In this context, a higher type frequency corresponds to the suffix appearing with a larger number of different stems relative to the competing suffix, while a larger token frequency corresponds to simply appearing a greater number of times relative to the competing suffx, regardless of the number of different stems it occurs with.

Procedure

Each participant was randomly assigned one of the conditions. In each condition, participants were first presented with an exposure phase. After the exposure phase, participants were tested using a production task, a form choice task, and a comprehension task. We describe each of these below.[1]

Exposure Phase

[1]Additionally, a demo of the experiment can be found at the following link: https://run.pavlovia.org/znhoughton/generalizability_demo.

Following Harmon & Kapatsinski (2017), each exposure trial consisted of the presentation of a picture on the computer screen which was subsequently followed with a written label for the image as well as an audio presentation of that label. Specifically, the image first appeared on the screen and then 1.25 seconds later was followed by both the label of the creatures on the screen as well as the audio for that creature. Participants were instructed to type the name of the creature and press enter. Participants had 4 seconds to respond after the presentation of the name of the creature and were given feedback as to whether they were correct or not.

Participants saw each trial within the exposure phase 5 times, each in a randomized order.

Production Task

After the exposure phase, participants were presented with a production task. In this task, participants were presented with images and told to produce a label for the image. Specifically, initially the unafixxed form appeared on the screen along with the corresponding image. 2 seconds later, the four different possible images of that creature appeared (a singular big creature, multiple big creatures, a single small creature, and multiple small creatures). Three of these images disappeared after 1.25 seconds, leaving a single image for the participant to produce a label for. Participants had 10 seconds to respond. In the production task, some of the stems were familiar (i.e., seen in training) and others were novel (i.e., not seen in training).

Participants saw each trial within the production task 4 times, each in a randomized order.

Form Choice Task

After the production task, participants were presented with a form choice task. In this task, participants were presented first with the base, unaffixed form and the corresponding image. 2 seconds later four images flashed on the screen, remained on the screen for 1.25 seconds, and disappeared leaving a single image. Along with the single image, participants were also presented with two possible labels for that image, one label on the bottom right of the screen and one label on the bottom left of the

screen. Participants were given four seconds to press either the left arrow or the right arrow to choose the corresponding label. The labels were counterbalanced with respect to which side of the screen they appeared on. In the form choice task, some of the stems were familiar (i.e., seen in training) and others were novel (i.e., not seen in training). The goal of this task was to assess whether type and/or token frequency influence the form choice when accessibility differences between frequent and rare forms have been attenuated.

Participants saw each trial within the form choice task 2 times, each in a randomized order.

Comprehension Task

Finally, participants were presented with a comprehension task. In this task, participants were first given the label and corresponding audio for a given creature. After 0.25 seconds, four images appeared on the screen and participants had 4 seconds to click one of the images on the screen that corresponded to the label. Similar to the before-mentioned tasks, in the comprehension task some of the stems were familiar (i.e., seen in training) and others were novel (i.e., not seen in training).

Participants saw each trial within the comprehension task 2 times, each in a randomized order.

Analyses and Results

Production Task

In order to examine the effects of type and token frequency on participants choice of suffix, we examined the effects of type and token frequency on the original meaning as well as the novel meaning (diminutive plural).[2]

[2]All data and code for the analyses can be found here: https://github.com/znhoughton/Generalizability-Type-Token.

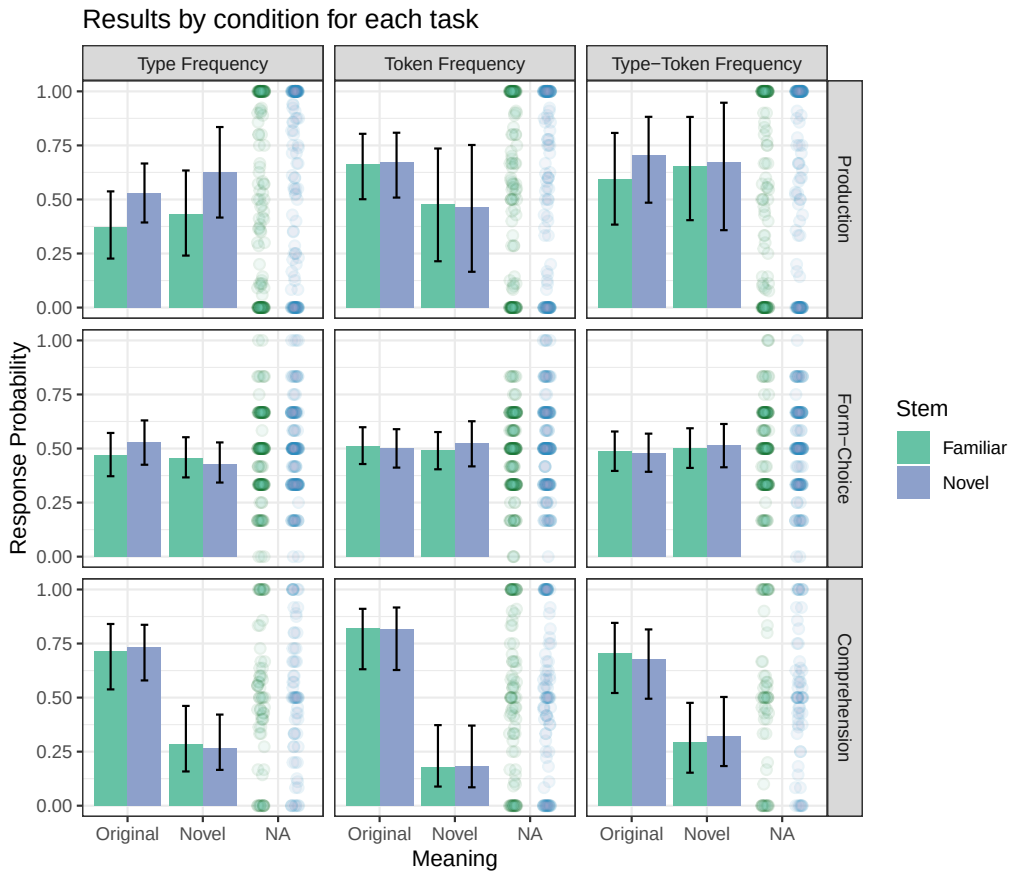

Figure 2: Plot of our results. Points indicate individual subject values. The x-axis indicates whether the meaning was original or novel. Green coloring corresponds to the frequent suffix while blue corresponds to the infrequent suffix. The y-axis indicates the response probability. For Production and Form-choice, it is the response probability of choosing a form. For Comprehension, it is the response probability of choosing a meaning. Facets indicate condition (type frequency, token frequency, or type-token frequency) and task (production, form-choice, comprehension).

In order to determine whether the effect of frequency on semantic extension differed between conditions, we ran a Bayesian linear mixed-effects model on the production data. The dependent variable was whether the participant produced the frequent suffix (*dan* in the *dan* language, *nem* in the *nem* language). A larger Intercept indicates that the frequent suffix was more likely to be produced than the infrequent suffix. A larger coefficient estimate indicates that the frequent form was more likely to be produced relative to the type-token frequency condition. The independent variables were condition, meaning, and stem (either familiar stem or novel stem). Condition was treatment coded such that the intercept was the high type and high token frequency condition. Meaning was sum coded. We also included a random intercept for participant and a random slope for meaning by participant. The syntax for our model is included below in Equation 1.

$$\text{frequent\_suffix} \sim \text{condition} * \text{meaning} + (1 + \text{meaning}|\text{participant}) \tag{1}$$

The results are shown in Table 2. The results suggest that in general the frequent suffix is used more than the infrequent suffix for all conditions. Notably however, we find no meaningful interaction effect for type frequency and novel meaning, suggesting that the effect of meaning is similar for both conditions in which a suffix had higher type frequency when it was frequent. We do, however, find an interaction effect between token frequency and novel meaning, suggesting that the frequent suffix is used less for the novel meaning when it has higher token frequency but the same type frequency as its competitors.

These results suggest that participants are more likely to produce the frequent suffix to convey a novel meaning when the suffix has both high type and token frequency, but not when there is high token frequency or high type frequency alone. Further, when a learner is expressing a familiar meaning, a high token frequency increases the probability of the learner producing the more frequent suffix.

In order to follow up on this, we ran an additional model that included whether the stem was

Table 2: Results of the statistical models for the production task.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	1.01	0.56	-0.05	2.13	96.87
Type Frequency	-0.75	0.71	-2.22	0.58	13.79
Token Frequency	-0.20	0.68	-1.57	1.13	38.40
Novel Meaning	0.10	0.32	-0.54	0.74	62.06
Type Frequency:Novel	0.25	0.42	-0.57	1.08	72.89
Token Frequency:Novel	-0.59	0.44	-1.47	0.26	8.69

familiar (occurred in training) or novel (did not occur in training). The results are reported in Table 3 and visualized in Figure 3.

The results demonstrate that there is little effect of stem familiarity for the type-token frequency and token frequency conditions. However, there is a pretty large effect of stem for the type frequency condition. Specifically, learners are more likely to use the frequent suffix in the type frequency condition when the stem is novel, and even more likely to use the frequent suffix when both the stem and meaning are novel.

Overall, our results suggest that in production, increased token frequency leads to learners choosing the frequent suffix to express the original meaning more than its competitors. Increased type frequency leads to learners choosing the frequent suffix to express the novel meaning more. Increasing both type and token frequency leads to learners choosing the frequent suffix to express both novel and familiar meanings more.

Form Choice Task

In order to examine the effects of form-choice, we similarly ran a model analogous to the one we ran for the production task (Equation 1). In the context of the form-choice task, the dependent variable reflects whether participants chose the option with the frequent suffix or the one with the infrequent suffix.

The results of the form-choice task suggest that when participants are presented with both

Table 3: Results of the statistical models for the production task.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	0.74	0.40	-0.04	1.54	96.93
Type Frequency	0.86	0.62	-0.35	2.10	92.10
Token Frequency	-0.69	0.58	-1.83	0.44	11.29
Novel Meaning	0.00	0.24	-0.46	0.49	48.90
Novel Stem	0.29	0.10	0.10	0.49	99.64
Type Frequency:Novel Meaning	0.36	0.36	-0.34	1.08	84.58
Token Frequency:Novel Meaning	0.31	0.33	-0.33	0.95	83.83
Type Frequency:Novel Stem	-0.02	0.12	-0.25	0.21	43.86
Token Frequency:Novel Stem	0.35	0.11	0.14	0.56	99.89
Novel Meaning:Novel Stem	0.09	0.08	-0.07	0.26	86.67
Type Frequency:Novel Meaning:Novel Stem	-0.10	0.11	-0.32	0.13	18.58
Token Frequency:Novel Meaning:Novel Stem	0.26	0.10	0.06	0.47	99.34

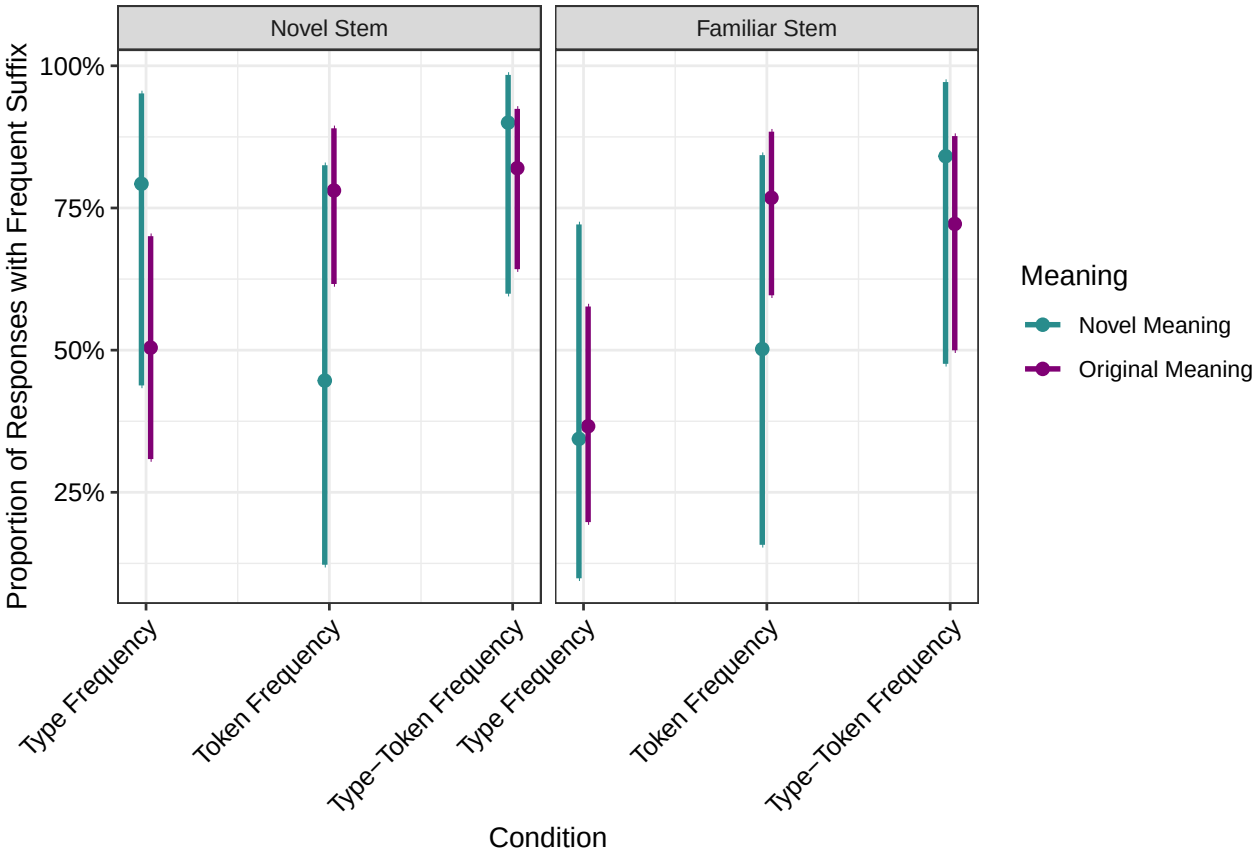

Figure 3: Plot of the statistical model estimates for the production task. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while purple points corresponds to the original meaning. The facet indicates whether the stem was novel or familiar.

Table 4: Results of the statistical models for the 2afc task.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	-0.01	0.07	-0.15	0.12	40.99
Type Frequency	-0.12	0.10	-0.31	0.07	10.31
Token Frequency	0.05	0.09	-0.13	0.22	70.61
Novel Meaning	0.06	0.07	-0.08	0.19	79.22
Type Frequency:Novel	-0.18	0.09	-0.36	0.00	2.50
Token Frequency:Novel	-0.05	0.09	-0.22	0.12	30.24

possible options, for both token frequency and type-token frequency, participants are equally likely to choose the frequent suffix as the infrequent suffix (i.e., token and type frequency don't matter) Interestingly, for the type frequency condition, participants were less likely to choose the frequent suffix for the novel meaning than the infrequent suffix, though the effect size is quite small and their selections were still close to chance. We also ran an additional model with stem condition as a fixed-effect (along with its interaction with condition and meaning) and found no effect of stem condition.

Comprehension Task

Our results for the comprehension task are included below in Table 5. For the comprehension task, the dependent variable for our statistical analysis was whether the meaning that participants chose was novel or original. A positive estimate suggests that participants were more likely to choose the novel meaning, a negative estimate suggests that participants were more likely to choose the original meaning.

For comprehension, our independent variables were different from the other tasks. Specifically, to maximize our power, we divided the data up based on whether the token-to-type ratio for the form presented on a given trial was 12-12, 12-3, or 3-3. For example, in the Type Frequency condition, while the frequent suffix occurs 12 times with 12 different types, the infrequent occurs 12 times in only 3 different types. By comparing the 12-3 condition to the 12-12 and 3-3 conditions, we can see how an increase in type frequency and a decrease in token frequency affect participants' choice of the novel or original meaning.

Table 5: Results of the statistical model for the comprehension task.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (12-3)	-2.80	0.27	-3.35	-2.28	0
12-12	1.46	0.19	1.09	1.84	100
3-3	1.98	0.18	1.62	2.34	100

We treatment coded condition such that the intercept was the 12-3 condition. As such, for each condition, a positive coefficient indicates that participants are more likely to choose a novel meaning (relative to the 12-3 condition/intercept). The model syntax is included below in Equation 2.

$$\text{meaning} \sim \text{condition\_numeric} + (1|\text{participant}) \tag{2}$$

First, we find a meaningful effect for the intercept (12-3 condition), suggesting that participants are more likely to select the original meaning when there is a high token-to-type ratio. Additionally, we find positive coefficient estimates for both 12-12 and 3-3 suggesting that when there is an equal number of tokens as types, participants are more likely to select the novel meaning. In other words, entrenchment seems to be driven not simply by the raw value of token or type frequency, but rather the proportion of tokens to types. That is, it is the relationship between type frequency and token frequency that is relevant for semantic entrenchment in comprehension.

Similar to the form-choice task, we also ran a model with stem condition as a main-effect as well as an interaction effect with condition and found no main-effect or interaction effect.

This mirrors what we saw in the results for our production task where having a high type and high token frequency leads to semantic extension, but having a high token frequency with low type frequency leads to entrenchment.

Rescorla-Wagner Model

In this section we examine whether the Rescorla-Wagner model (Rescorla, 1972), an associative learning model, can accurately predict our data or not.

The Rescorla-Wagner model is an associative learning model that has gained a great deal of popularity. Despite its simplicity, it has been shown to account for many phenomena in language learning and processing (Baayen, 2012; Ellis, 2006; Olejarczuk et al., 2018; Ramscar et al., 2013).

Specifically, the Rescorla-Wagner model is a two-layer feedforward neural network. It takes as its input a set of cues and predicts an outcome (Kapatsinski, 2021). The Rescorla-Wagner model is defined below in Equation 3 and Equation 4, which differ in whether the predicted outcome is present or not. The model learns the weight of an association between a present cue and a present outcome. In the equation, α_C denotes the salience of the cue, β_O^P denotes the salience of an outcome when it is present, and β_O^A denotes the salience of an outcome when it is absent (Kapatsinski, 2021). λ_{max} is the learning rate.

$$\Delta V_{C \to O} = \alpha_C \beta_O^P (\lambda_{max} - \alpha_O) \tag{3}$$

$$\Delta V_{C \to O} = \alpha_C \beta_O^A (\lambda_{min} - \alpha_O) \tag{4}$$

The Rescorla-Wagner model originally uses a linear activation function, however for categorical outcomes a logistic activation performs better because it avoids the prediction of spurious excitement, where a cue becomes associated with an outcome even if never occurring with it (Kapatsinski, 2021). Following the suggestion in Kapatsinski (2021), we use a logistic activation function (Equation 5) for the simulations in this section.

$$\alpha_O = logit^{-1}(\sum_c V_{C \to O}) \tag{5}$$

Given the Rescorla-Wagner model's success on a wide variety of linguistic phenomena, it's possible that it is also able to predict the results we see in this paper. Thus in this section we simulate the Rescorla-Wagner predictions for our production data in each exposure condition and compare it to the human data.

For the simulations in this section, α and β were set to 0.1, and λ was set to 2. Our simulation process was as follows: First, for each of the different exposure conditions, we obtained a matrix of predicted cue-outcome associations using the Rescorla-Wagner model. Then for each of our items in the production task, we summed the associations for each cue-outcome present. For example, given the item *baldan* and the meaning big.pl, we consulted the matrix of cue-outcome associations for *bal* and *dan*, big and *dan*, and pl and *dan*. We then summed these to get the predicted associations for *bal* with the big.pl meaning and *dan*.

After calculating these cue-outcome associations for each test item, we took the difference between the cue-outcome association for the frequent suffix and the cue-outcome association for the infrequent suffix. Thus for each human response, we now also had the Rescorla-Wagner activation difference between the frequent and infrequent suffix. If the associations are accurate, then the difference between these values should be a good predictor of whether humans chose the frequent suffix or the infrequent suffix on each trial.

In order to test our model predictions, we created three Bayesian models to determine which model predicted the data best. The first model was a simple model that modeled frequency as a function of the Rescorla-Wagner predictions with a random intercept for participant and for stem (Equation 6). The second model was analogous, except that we divided the activations into stem and meaning activations. For example, instead of having a single activation for `bal.dim.sg` (which was the association between those cues and the frequent suffix minus the association between those cues and the infre-

quent suffix) we instead calculated the activations separately for the stem, *bal*, and the meaning, dim.sg (Equation 7). Finally, in our third statistical model, we included the original predictors in the human model (stem, condition, and meaning) along with the Rescorla-Wagner predictions (Equation 8). In all of our models, frequency is whether the frequent or infrequent suffix was chosen in the human data, meaning is whether the meaning was familiar or novel, stem_condition was whether the stem was familiar or novel, freq_minus_infreq was the difference in predicted associations between the frequent and infrequent suffixes, and resp_stem was the individual stem (e.g., *bal*).

$$\text{frequency} \sim \text{freq\_minus\_infreq} + (\text{freq\_minus\_infreq}|\text{participant}) + (1|\text{resp\_stem}) \tag{6}$$

$$\begin{aligned}\text{frequency} \sim\ & \text{freq\_minus\_infreq\_stem} * \text{freq\_minus\_infreq\_meaning} \\ & + (\text{freq\_minus\_infreq\_stem} * \text{freq\_minus\_infreq\_meaning}|\text{participant}) \\ & + (1|\text{resp\_stem})\end{aligned} \tag{7}$$

$$\begin{aligned}\text{frequency} \sim\ & \text{condition} * \text{meaning} * \text{stem\_condition} + \text{freq\_minus\_infreq} \\ & + (1 + \text{stem\_condition} * \mathit{meaning} + \text{freq\_minus\_infreq}|\text{participant}) \\ & + (1|\text{resp\_stem})\end{aligned} \tag{8}$$

In order to compare which model best fit the data, we performed leave-one-out cross-validation using the R package `loo` (Vehtari et al., 2017). Leave-one-out cross-validation refits the model for each observation in the dataset, leaving out that observation.[3] For each data point, the expected log predictive density is calculated. Expected log predictive density is the log-likelihood

[3] More accurately, Bayesian models are computationally expensive to fit many times, so this is approximated using Pareto-smoothed importance sampling instead.

Table 6: Results of our leave-one-out cross-validation.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Stem and Suffix Activations as Separate Predictors	0.00	0.00	-278.90	19.05
Original Predictiors Plus RW Activations	-40.34	7.50	-319.24	20.53
Only RW Activations	-110.81	14.12	-389.71	24.81
Only Original Predictors	-1286.41	28.12	-1565.31	26.01

Table 7: Results of the statistical analysis containing the predictions from the Rescorla-Wagner logistic model separated into stem activations and meaning activations. The results demonstrate that the human data is predicted well by suffix activations but not by stem activations.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-4.45	1.03	-6.74	-2.69	0.00
Stem Activations	0.67	0.48	-0.31	1.59	91.50
Suffix Activations	11.11	1.75	8.28	15.12	100.00
Stem:Suffix Activations	-0.01	0.86	-1.74	1.76	49.27

of the held-out observation given the model's parameters (trained without the observation). Each model receives a single expected log predictive density value by summing the log-likelihood of each observation using leave-one-out cross-validation. A higher value indicates the model performs better than the other models. Thus, we compared the expected log predictive density value for each of these three models, along with the model of the human data that does not contain the Rescorla-Wagner predictions.

The results of our leave-one-out cross-validation are included in Table 6. Our results suggest that in general, every model that includes the Rescorla-Wagner predictions outperforms the model without them. Interestingly, the best fitting model is the one that separates stem activations and meaning activations into separate predictors. The results of this model are presented in Table 7 and suggest that the human data is predicted well by the meaning activations. The model also shows little effect of stem activations (although over 90% of the posterior samples were greater than zero), however a follow-up model containing only stem activations (and not meaning activations) demonstrated a meaningful effect of stem activations. Thus, the non-significant effect of stem activation in Table 7 is more likely due to colinearity than it is a lack of effect of stem activations.

The results of our simulations suggest that an error-driven associative learning model, specifically the Rescorla-Wagner model, does a good job of fitting the data. Interestingly, a model with only the activations outperforms (by a large margin) the model that includes stem, condition, and meaning as predictors. Altogether, these simulations demonstrate that an associative learning model accurately predicts the effects of token and type frequency on semantic extension.

Discussion

Our results suggest that in production, having a high type and high token frequency (relative to the competitor), along with simply having a high type but equal token frequency (relative to the competitor), leads to semantic extension, while having a high type frequency but low token frequency leads to entrenchment (i.e., using the suffix more with the original meaning than novel meanings). Further, this mirrors what we see in comprehension where participants are more likely to choose the novel meaning when the type and token frequencies are matched, but less likely when the suffix has high token frequency but low type frequency. Further, when both forms are made accessible as in the form choice task, the preference for the frequent suffix over the infrequent suffix disappears for all conditions.

Our results also demonstrate that the Rescorla-Wagner model with a logistic activation function does a good job of capturing the human data. This suggests that the effects we see are explained by an associative learning model.

Baayen, H. (2012). *Demythologizing the word frequency effect: A discriminative learning perspective* (G. Libben, G. Jarema, & C. Westbury, Eds.; Vol. 47, pp. 171–195). John Benjamins Publishing Company. https://doi.org/10.1075/bct.47.10baa

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44. https://doi.org/10.1016/j.cogpsych.2017.08.002

Kapatsinski, V. (2021). Learning fast while avoiding spurious excitement and overcoming cue competition requires setting unachievable goals: Reasons for using the logistic activation function in learning to predict categorical outcomes. *Language, Cognition and Neuroscience*, *0*(0), 1–22. https://doi.org/10.1080/23273798.2021.1927120

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*(s2), 1–9. https://doi.org/10.1515/lingvan-2017-0020

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017–1023. https://doi.org/10.1177/0956797612460691

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory*, *2*, 6469. https://cir.nii.ac.jp/crid/1572543025504096640

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4