The effects of type and token frequency on semantic extension

Introduction

Methods

Following Harmon & Kapatsinski (2017), two artificial languages were used: called 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 can examine how properties of a suffix distribution (type frequency and token frequency) affect its likelihood of being extended to express the diminutive plural meaning.

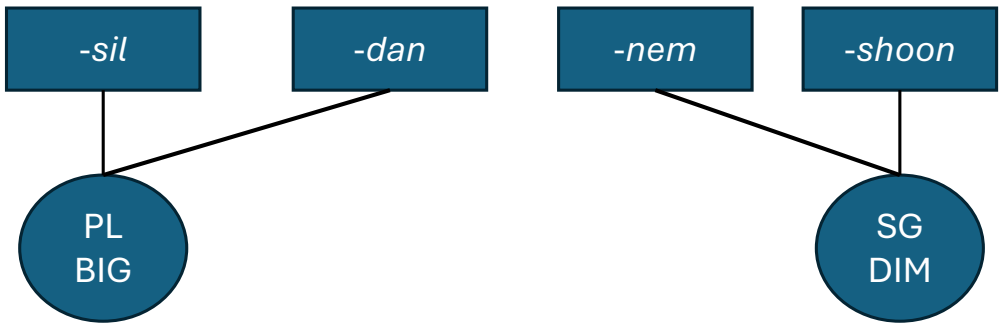

Figure 1: A description of the suffixes in our artificial languages. The thicker lines denote the more frequent form in each language: the plural *-dan*$_{PL}$ in the Dan language and the diminutive *-nem*$_{DIM}$ in the Nem language.

During the exposure phase, each suffix was 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 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 suffix (i.e., *dan* in Dan and *nem* in Nem) was higher than the type frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3) while the token frequency of the two suffixes was matched and set equal to the type frequency of the frequent suffix (both at 12). We refer to this condition as Type. In the second condition, the token frequency of the frequent suffix was higher than the token frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3) while holding the type frequency of the two suffixes equal to the type frequency of the infrequent suffix (both at 3). We refer to this condition as Token. Finally, in the third condition, we manipulated both the type and token frequency of the frequent suffix such that the token and type frequency of the frequent suffix were both greater than the token and type frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3). We refer to this condition as Type-Token. In this context, a higher type frequency corresponds to the suffix appearing with a larger number of distinct stems relative to the competing suffix, while a larger token frequency corresponds to simply appearing a greater number of times relative to the competing suffix, regardless of the number of different stems it occurs with.

Our Experiment comprised of three different conditions (see Table 1), varying type and token frequencies of the competing suffixes. Type frequency corresponds to the number of distinct stems with which a suffix appears, while token frequency corresponds to the number of occurrences of a suffix, regardless of the number of different stems it occurs with. In the Type condition, the type frequency of the frequent suffix (i.e., *dan* in Dan and *nem* in Nem) was higher than the type frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3) while the token frequency of the two suffixes was matched and set equal to the type frequency of the frequent suffix (both at 12). The frequent suffix in this condition has a lower token/type ratio than the rare suffix.

In the Token condition, the token frequency of the frequent suffix was higher than the token

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 Suffix		Infrequent Suffix	
Condition	**Type**	**Token**	**Type**	**Token**
Type	12	12	3	12
Token	3	12	3	3
Type-Token	12	12	3	3

frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3) while the type frequencies of the two suffixes were equal (both at 3). In this condition, both suffixes co-occurred with the same types. The frequent suffix in this condition has a higher token/type ratio than the rare suffix.

In the Type-Token condition, we manipulated both the type and token frequency of the frequent suffix such that the token and type frequency of the frequent suffix were both greater than the token and type frequency of the infrequent suffix (by a factor of 4 or 12 vs. 3). In this condition, the frequent and infrequent suffixes have the same token/type ratio.

Procedure

175 participants were recruited from Prolific, a crowd sourcing platform. 42 participants were excluded for failing to complete the task properly or achieving below 75% accuracy on control items. 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. 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]

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

Exposure Phase

Following Harmon & Kapatsinski (2017), each exposure trial consisted of the presentation of a picture on the computer screen which was followed by 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 (positioned directly below the image with a letter height equal to 0.05 times the height of the user's screen) as well as the audio corresponding to the label (22050 Hz, 16 bit) timed to start simultaneously. The label remained on the screen for 3 seconds and the image remained on the screen for the duration of the label. Participants were instructed to type the name of the creature and press Enter. Participants had 4 seconds to respond after the onset of the presentation of the name of the creature and were given feedback as to whether they were correct immediately after pressing Enter.

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 unaffixed form appeared on the screen along with the corresponding image. Two seconds later, the label disappeared and 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 text label for the base, unaffixed form and the corresponding image. Two seconds later, the label and the image disappeared and four images flashed on the screen, remained on the screen for 1.25 seconds before disappearing, leaving a single image. Along with the single image, participants were also presented with two possible labels (which differed in whether the suffix was *dan* or *nem*) 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, the label remained on the screen and four images appeared on the screen. Participants had 4 seconds to click one of the images on the screen that corresponded to the label. Similar to the aforementioned 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

We explore the results for each task in depth in the next sections.[2]

Production Task

In order to determine whether the effect of frequency on semantic extension differed between conditions, we ran a Bayesian logistic mixed-effects model on the production data. The dependent variable was whether the participant produced the frequent suffix. The frequent suffix was coded as 1 if participants chose *dan* in the Dan language or chose *nem* in the Nem language.

We treatment coded condition such that the intercept was the type-token condition. Thus, a larger intercept indicates that the frequent suffix was more likely to be produced than the infrequent suffix when it had a higher type and token frequency.

A larger coefficient estimate for a condition indicates that the frequent form was more likely to be produced than the infrequent form in that condition relative to the type-token frequency condition.

We also included meaning and stem novelty as sum-coded variables. Meaning was a categorical variable with two levels: original and novel. An original meaning referred to big plural if the suffix was *dan* or diminutive singular if the suffix was *nem*. A meaning was novel if it was diminutive plural regardless of the suffix. Stem novelty similarly was a categorical variable with two levels, familiar or novel. A familiar stem was one that participants saw in training while a novel stem was one that they did not see in training. We also included a random intercept for participant and random slopes for meaning by participant and stem novelty by participant. The syntax for our model is included below in Equation 1. The results are reported in Table 2 and visualized in Figure 2 and Figure 3.

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

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

We find a meaningful effect for the intercept, suggesting that the frequent suffix is chosen more than the infrequent suffix when it was higher in both type and high token frequency.

We also find a meaningful effect for novel stems, suggesting that the effect of suffix frequency is greater for novel stems than for familiar stems in the Type-Token condition. The effect of stem novelty is even greater in the Type condition, where there is a preference for the frequent suffix over the infrequent suffix when the stem is novel but not when it is familiar (evidenced by the interaction effect between type frequency and novel stem). In contrast, there is no suffix-frequency-by-stem-novelty interaction in the Token condition.

Finally, we find a three-way interaction between Type frequency condition, novel meaning, and novel stem, suggesting that learners are especially likely to produce a frequent suffix in the Type frequency condition when both the stem and meaning are novel.

Overall, our results suggest that in production, learners generally use the frequent suffix more regardless of meaning or stem familiarity if the suffix has both a high type and high token frequency. When the frequent suffix has a higher type frequency than the infrequent suffix, learners are especially likely to use it if the stem is novel. When the stem is familiar, the frequent suffix is preferred over the infrequent one only if they differ in token frequency. Finally, a frequent suffix is preferred over the infrequent suffix in the novel meaning only if they differ in type frequency.

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

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

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	1.02	0.55	-0.05	2.14	96.91
Type Frequency	-0.74	0.71	-2.20	0.58	14.57
Token Frequency	-0.22	0.68	-1.59	1.12	37.24
Novel Meaning	0.09	0.33	-0.54	0.74	61.03
Novel Stem	0.27	0.14	-0.01	0.56	97.06
Type Frequency:Novel Meaning	0.28	0.42	-0.55	1.12	74.46
Token Frequency:Novel Meaning	-0.61	0.45	-1.51	0.26	8.42
Type Frequency:Novel Stem	0.36	0.17	0.02	0.70	98.11
Token Frequency:Novel Stem	-0.29	0.21	-0.70	0.12	7.86
Novel Meaning:Novel Stem	-0.01	0.13	-0.26	0.25	46.68
Type Frequency:Novel Meaning:Novel Stem	0.36	0.17	0.03	0.68	98.34
Token Frequency:Novel Meaning:Novel Stem	-0.05	0.20	-0.44	0.35	40.76

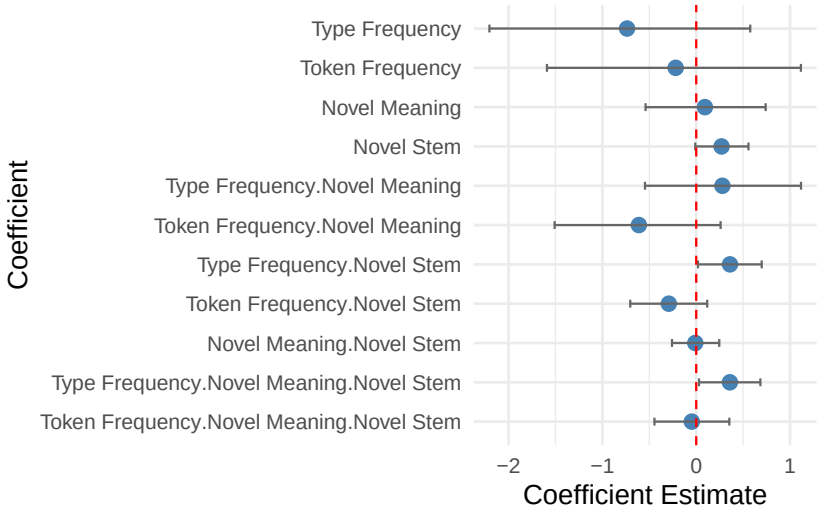

Figure 2: Plot of the coefficient values in the statistical model presented in Table 2. The x-axis indicates the coefficient estimate while the y-axis indicates the different coefficients in the model. Blue points indicate the posterior mean estimate for each coefficient while the gray bars indicate the 95/% credible interval for each estimate. The intercept represents the baseline condition (Type-Token Frequency). All other coefficients are expressed as changes relative to this baseline, so the intercept is omitted from the plot.

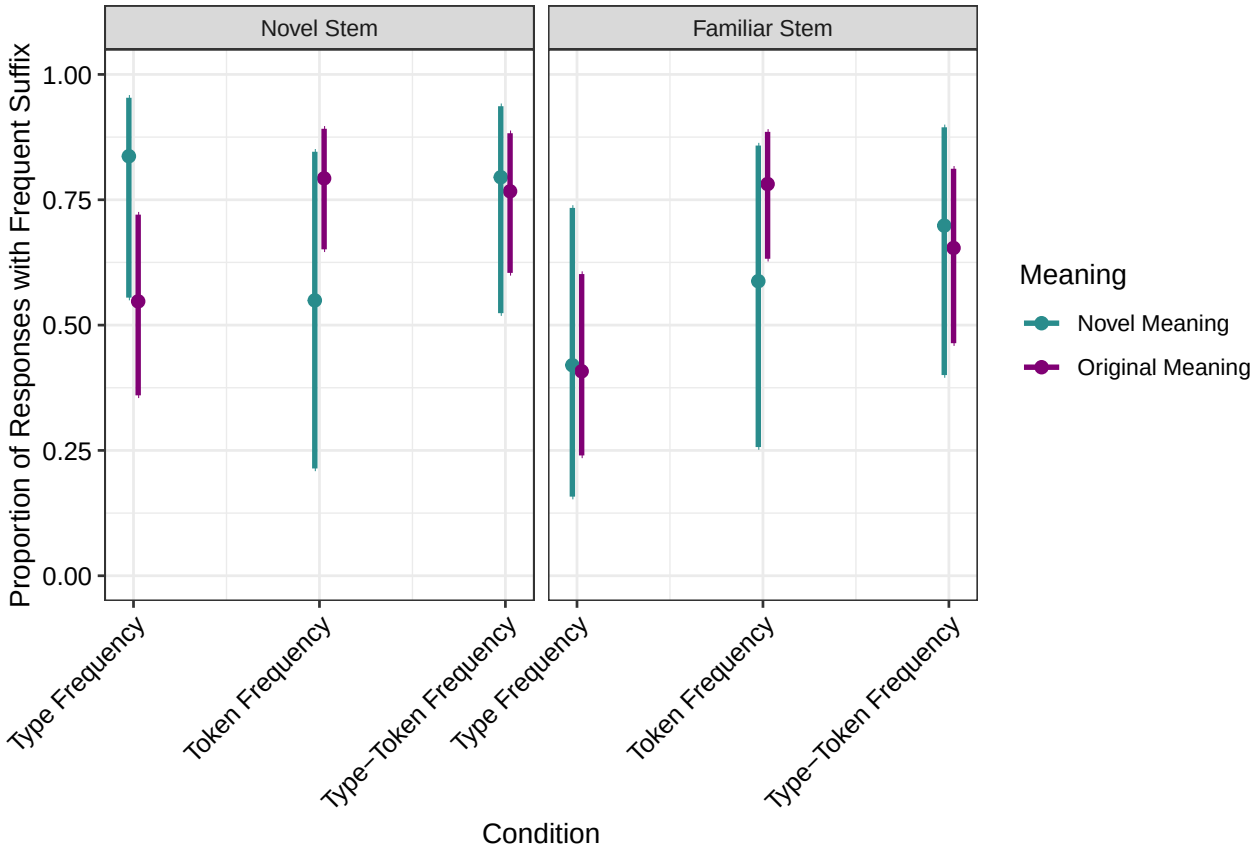

Figure 3: Plot of the statistical model estimates for the production task. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while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

Table 3: Results of the statistical models for the form-choice task.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	-0.01	0.07	-0.15	0.12	41.23
Type Frequency	-0.12	0.10	-0.30	0.06	10.50
Token Frequency	0.05	0.09	-0.13	0.22	70.71
Novel Meaning	0.05	0.07	-0.08	0.19	78.69
Novel Stem	-0.01	0.07	-0.14	0.12	45.84
Type Frequency:Novel Meaning	-0.18	0.09	-0.36	0.00	2.81
Token Frequency:Novel Meaning	-0.05	0.09	-0.22	0.13	30.07
Type Frequency:Novel Stem	-0.03	0.09	-0.21	0.15	37.80
Token Frequency:Novel Stem	-0.02	0.09	-0.19	0.15	41.59
Novel Meaning:Novel Stem	-0.02	0.07	-0.15	0.11	38.23
Type Frequency:Novel Meaning:Novel Stem	0.12	0.09	-0.07	0.30	89.33
Token Frequency:Novel Meaning:Novel Stem	-0.03	0.09	-0.20	0.14	35.20

infrequent suffix. The model was exactly the same as in production. The results are presented in Table 3 and visualized in Figure 4 and Figure 5.

In general we find no reliable effects except for an interaction effect between type frequency and novel meaning, suggesting that there may be a preference for the frequent suffix when it has a higher type frequency than the infrequent suffix. However, the small size of this effect and the lack of any other meaningful effects suggests that when participants are presented with both possible options, for all conditions (except when there is a novel meaning in the type frequency condition), participants are equally likely to choose the frequent suffix as the infrequent suffix (i.e.,token and type frequency don't seem to effect this).

Finally, in order to verify that the differences in the results between form-choice and production are meaningful, we ran a model with task (form-choice vs production) as an interaction effect. The results are presented in Table 4 and visualized in Figure 6. The results demonstrate that learners are overall more likely to select the frequent suffix in the production task than in the form-choice task. The results also confirm that there is an effect of stem familiarity in production, but this effect is greatly reduced in the form-choice task. Additionally, the results demonstrate that in production, there is an interaction effect between type frequency, novel stems, and novel meanings, such that learners are especially likely to select the frequent suffix when both the stem and meaning are novel. However, this

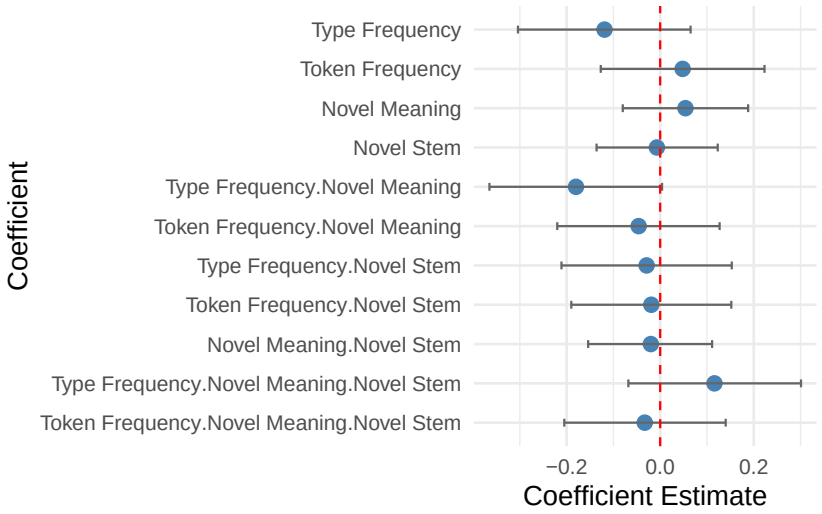

Figure 4: Plot of the coefficient values in the statistical model presented in Table 3. The x-axis indicates the coefficient estimate while the y-axis indicates the different coefficients in the model. Blue points indicate the posterior mean estimate for each coefficient while the gray bars indicate the 95/% credible interval for each estimate. The intercept represents the baseline condition (Type-Token Frequency). All other coefficients are expressed as changes relative to this baseline, so the intercept is omitted from the plot.

effect is diminished in the form-choice task (as demonstrated by the negative four-way interaction effect between type frequency, novel meaning, novel stem, and form-choice task). Finally, the results also demonstrate that in the token frequency condition, the effect of stem familiarity is greater in the form-choice task than in production however in the type frequency condition the opposite is found: the effect of stem familiarity is greater in the production task than in the form-choice task.

Comprehension Task

In order to examine the results for the comprehension task, similar to the previous tasks we ran a mixed-effects regression model. The dependent variable was whether the meaning that participants selected was novel or original. A positive estimate indicates that participants chose the novel meaning while a negative estimate indicates that participants were more likely to choose the original meaning. In production and form-choice, participants are choosing between suffixes that differ in frequency. In comprehension, participants instead are presented with a single affix that has a certain type frequency, token frequency, and token/type ratio, and a frequent suffix from one condition can have

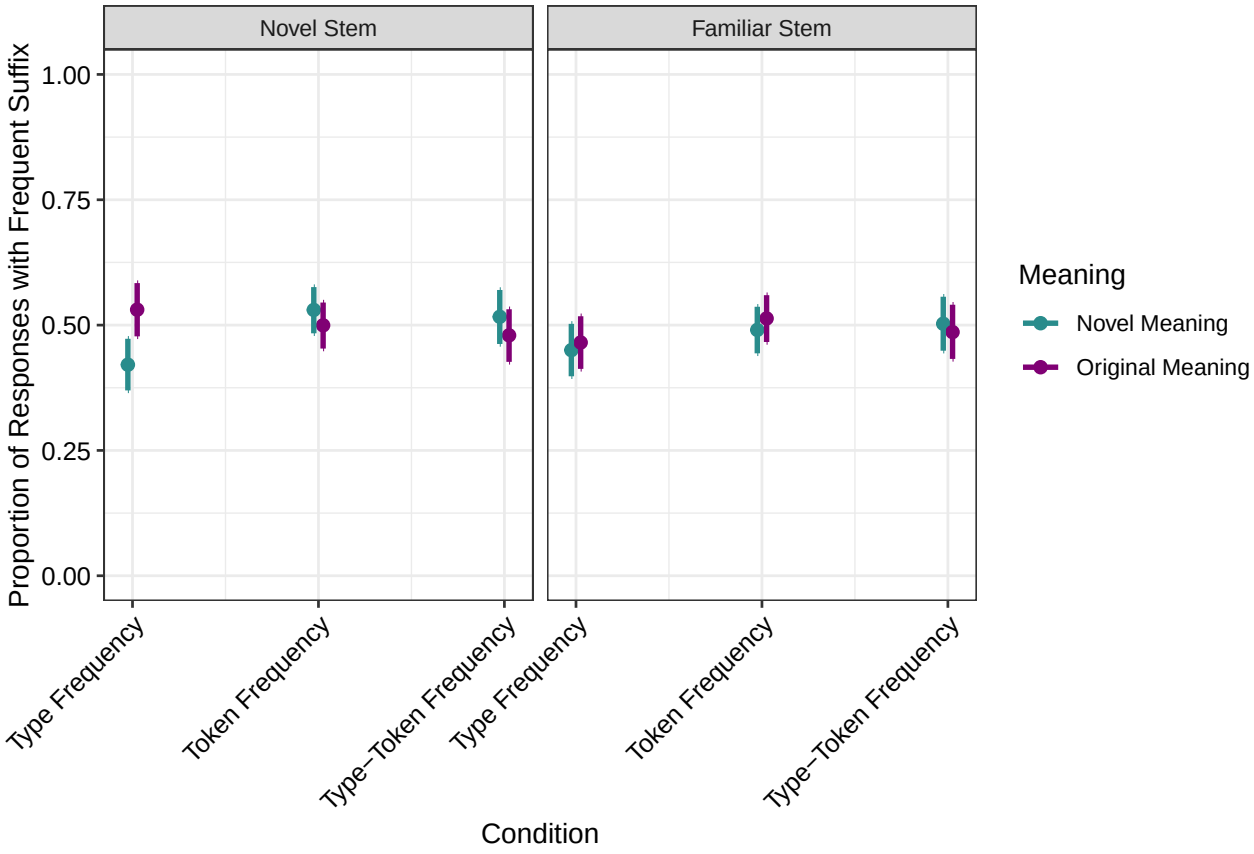

Figure 5: Plot of the statistical model estimates for the form-choice task. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while purple points indicate that the meaning was original. The facet indicates whether the stem was novel or familiar.

Table 4: Results of the statistical models with task (2afc vs production) as an interaction effect.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Type-Token Frequency)	0.36	0.21	-0.06	0.78	95.64
Type Frequency	-0.29	0.27	-0.84	0.22	13.27
Token Frequency	0.01	0.25	-0.49	0.52	51.59
Novel Meaning	0.03	0.13	-0.23	0.30	59.63
Novel Stem	0.13	0.06	0.00	0.26	97.92
Form-Choice	-0.38	0.21	-0.79	0.03	3.34
Type Frequency:Novel Meaning	0.06	0.17	-0.28	0.40	63.34
Token Frequency:Novel Meaning	-0.28	0.18	-0.66	0.07	5.75
Type Frequency:Novel Stem	0.19	0.09	0.01	0.36	98.35
Token Frequency:Novel Stem	-0.13	0.10	-0.33	0.06	8.85
Novel Meaning:Novel Stem	0.00	0.07	-0.12	0.13	51.14
Type Frequency:Form-Choice	0.16	0.26	-0.33	0.70	73.55
Token Frequency:Form-Choice	0.04	0.25	-0.46	0.53	56.25
Novel Meaning:Form-Choice	0.03	0.14	-0.24	0.29	58.86
Novel Stem:Form-Choice	-0.13	0.07	-0.26	0.00	2.63
Type Frequency:Novel Meaning:Novel Stem	0.12	0.09	-0.05	0.29	91.66
Token Frequency:Novel Meaning:Novel Stem	-0.01	0.10	-0.20	0.18	47.01
Type Frequency:Novel Meaning:Form-Choice	-0.25	0.18	-0.60	0.09	7.75
Token Frequency:Novel Meaning:Form-Choice	0.23	0.19	-0.13	0.60	89.39
Type Frequency:Novel Stem:Form-Choice	-0.16	0.09	-0.33	0.01	3.76
Token Frequency:Novel Stem:Form-Choice	0.15	0.10	-0.04	0.35	94.12
Novel Meaning:Novel Stem:Form-Choice	0.02	0.07	-0.11	0.15	62.26
Type Frequency:Novel Meaning:Novel Stem:Form-Choice	-0.24	0.09	-0.41	-0.07	0.30
Token Frequency:Novel Meaning:Novel Stem:Form-Choice	0.04	0.10	-0.15	0.23	65.69

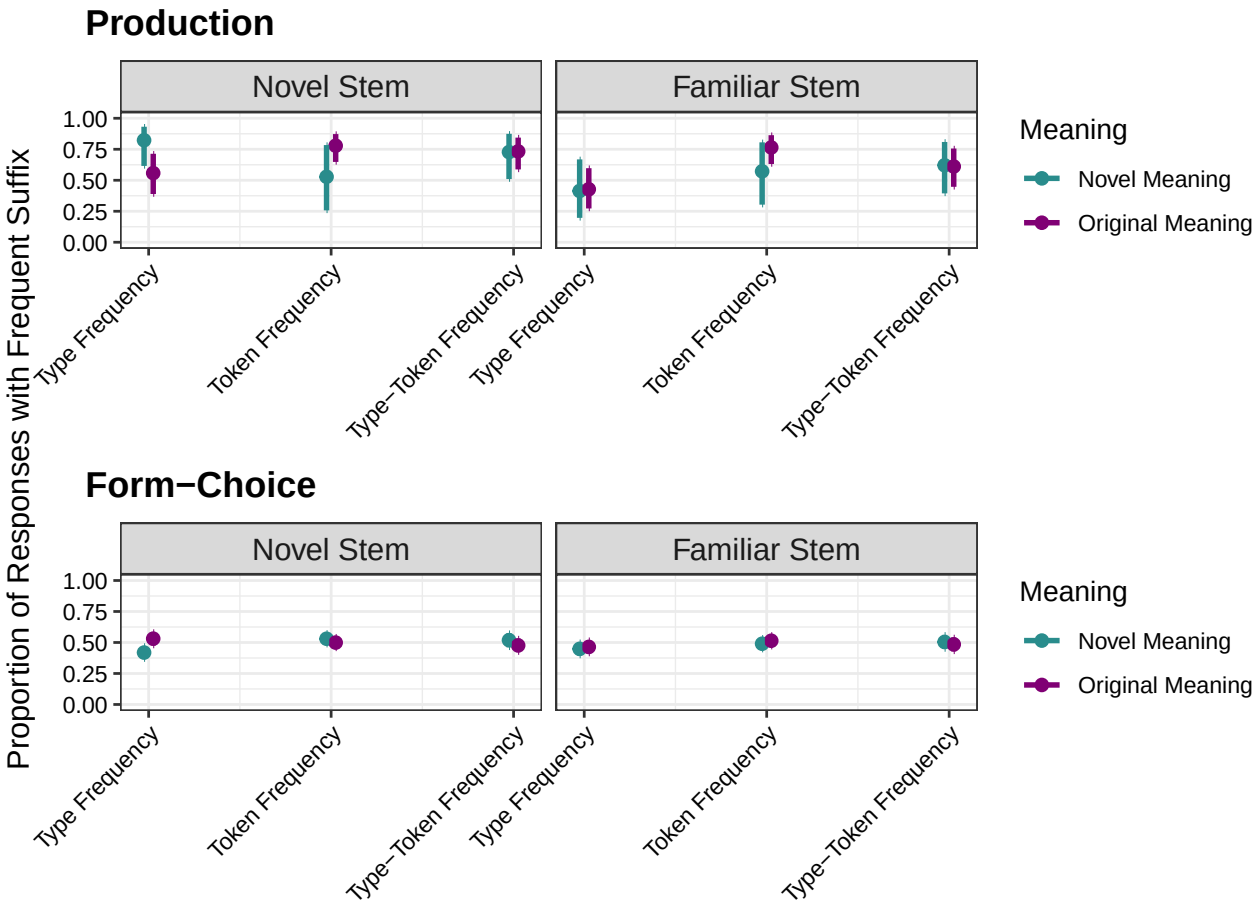

Figure 6: Plot of the statistical model estimates for task-interaction model. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while purple points indicate that the meaning was original. The facet indicates whether the stem was novel or familiar. The production results are shown in the top plot and the form-choice results are shown in the bottom plot.

the same characteristics as an infrequent suffix from another condition. Since participants are not comparing suffixes in comprehension, we expect equally distributed suffixes to have the same effect. Thus, the independent variable chosen was the distribution of the presented suffix: 12 tokens and 12 types, 12 tokens and 3 types, or 3 tokens and 3 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 vs. original meaning. Thus, hereafter we refer to the 12-3 distribution as the baseline distribution, the 12-12 distribution as Higher Type and the 3-3 distribution as Lower Token. Another way to classify the distributions is that the baseline distribution is a suffix that has a higher token/type ratio than the others (4 tokens per type vs 1 token per type).

We treatment coded distribution such that the intercept was the baseline condition. As a result, for each distribution, a positive coefficient indicates that participants are more likely to choose a novel meaning (relative to the baseline distribution/intercept). We also included a slope for stem familiarity, a random intercept for participant, and a random slope for stem familiarity by participant.

The model syntax is included below in Equation 2. Our results for the comprehension task are included in Table 5 and visualized in Figure 7 and Figure 8.

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

First, we find a meaningful effect for the baseline distribution (Intercept), suggesting that participants are more likely to select the original meaning than the novel meaning. Additionally, we find positive coefficient estimates for both Higher Type and Lower Token conditions, suggesting that decreasing the token-to-type ratio, whether by increasing type frequency or decreasing token frequency, makes participants more likely to select the novel meaning. Finally, we find no effect of stem familiarity. These results suggest that it may not be type or token frequency independently that drives semantic entrenchment, but rather the ratio between type and token frequency.

Table 5: Results of the statistical model for the comprehension task with stem as a fixed-effect.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	-2.81	0.28	-3.38	-2.28	0.00
12-12	1.47	0.19	1.09	1.84	100.00
3-3	1.98	0.18	1.62	2.35	100.00
Novel Stem	0.02	0.12	-0.21	0.25	57.66
12-12:Novel Stem	-0.03	0.16	-0.34	0.28	43.70
3-3:Novel Stem	-0.11	0.16	-0.41	0.20	25.02

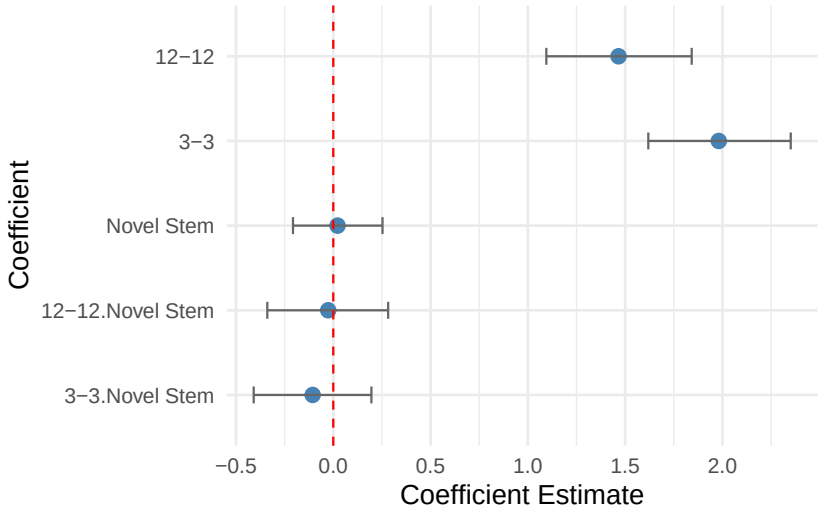

Figure 7: Plot of the coefficient values in the statistical model presented in Table 5. The x-axis indicates the coefficient estimate while the y-axis indicates the different coefficients in the model. Blue points indicate the posterior mean estimate for each coefficient while the gray bars indicate the 95/% credible interval for each estimate. The intercept represents the baseline condition (Type-Token Frequency). All other coefficients are expressed as changes relative to this baseline, so the intercept is omitted from the plot.

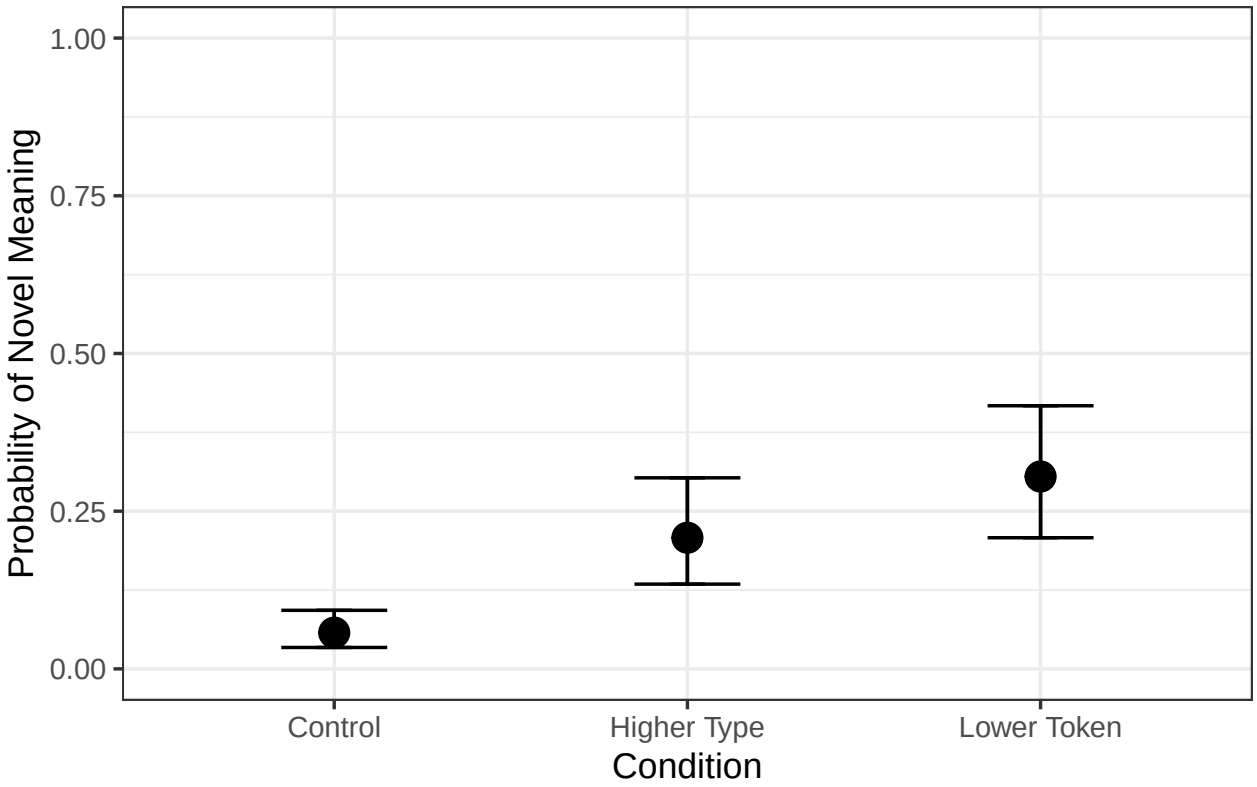

Figure 8: Plot of the model estimates for the comprehension task.

Overall, the comprehension results demonstrate that participants are more likely to select the original meaning when there is a high token-to-type ratio. However, when the token-to-type ratio is equivalent (either 12-12 or 3-3) then participants are more likely to select the novel meaning than when there is a high token-to-type ratio. One caveat is that participants still prefer the original meaning over the novel meaning in general (this is evident because adding the upper CI estimate to the intercept for both conditions still results in a negative estimate). In other words, while decreasing the token-to-type ratio decreases the preference for the original meaning, it does not eliminate it.

Connectionist Models

In this section we examine whether the results are accurately accounted for by a simple connectionist model implementing error-driven predictive learning; specifically, the logistic perceptron (Rumelhart & McClelland, 1986), which is an extension of the Rescorla-Wagner model to categorical

outcomes (Dawson, 2008; Kapatsinski, 2023). The logistic perceptron can be considered a standard model for learning to predict categorical outcomes and serves as the final layer in all modern neural network models. It is also the online counterpart to logistic regression. These models do not have a categorical distinction between types, but do generally adjust network beliefs more when a token from a novel type because a token from a novel type is generally less expected and learning is proportional to prediction error. Furthermore, type frequency helps generalization because it increases the likelihood that the generalization stimuli are within the similarity space covered by known stimuli (Hare et al., 1995). Therefore, much previous work has debated whether such models can capture effects of type frequency on morphological productivity, or if additional, symbolic mechanisms are needed to differentiate types (Caballero & Kapatsinski, 2022; del Prado Martın et al., 2004; Hare et al., 1995; Pinker & Prince, 1988; Xu & Tenenbaum, 2007).

The Rescorla-Wagner model is an error-driven predictive learning model that takes as its input a set of cues and predicts an outcome or set of outcomes as output (Rescorla & Wagner, 1972).

The 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 cues and outcomes ($V_{C\rightarrow O}$) . In the equation, α_C denotes the salience of a 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. Absent cues are generally assumed to have $\alpha_C = 0$ and are therefore not to be updated. Because we have no reason to believe some cues or outcomes are more salient than others, we assumed that $\alpha_C = 0.1$ for all present cues and $\beta = 1$ for both present and absent outcomes, yielding a learning rate of 0.1. The results do not qualitatively change with changes in learning rate. An outcome's activation, A_O, denotes the sum of the association weights (V) from the present cues to that outcome, i.e., $\sum_c V_{C\rightarrow O}$. Subtracting A_O from the appropriate limit (1 for present outcomes and 0 for absent ones) defines prediction error. The model is error-driven because the change in weight is proportional to prediction error: more is learned when the presence or absence of an outcome is surprising.

$$\Delta V_{C \to O} = \alpha_C \beta_O^P (1 - A_O) \tag{3}$$

$$\Delta V_{C \to O} = \alpha_C \beta_O^A (0 - A_O) \tag{4}$$

The logistic perceptron simply redefines prediction error as shown below, passing an outcome's activation through a logistic activation function before subtracting it from the limit (1 or 0; Equation 5). This ensures that the prediction error is always positive for present outcomes and negative for absent outcomes. The logistic perceptron is the standard way to predict categorical outcomes in neural networks (following Rumelhart et al., 1986). Furthermore, Caballero & Kapatsinski (2022) along with Kapatsinski (2023) show that this modification eliminates an incorrect prediction of the Rescorla-Wagner model, which otherwise sometimes learns associations between cues and outcomes that never co-occur, explains the transient nature of cue competition effects like blocking and overshadowing, and can account for S-shaped learning curves.

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

Production Models

In this section, we examine whether the logistic perceptron can account for the results of our production task.

The model was trained on the same series of training trials as our human subjects, presented in random order. The cues represented stem identity along with the meaning (e.g., bal_dim_pl) and the outcome was the suffix that the stem occurred with on that trial (i.e., *dan*, *nem*, *sil*, or *shoon*).

For each of the items in the production task, we computed activation of each outcome suffix. For example, given the item *baldan* and the meaning BIG PLURAL, we summed the weights of

associations from *bal*, BIG, and PLURAL to *dan*. The activation of *dan* corresponds to its expected log odds given these cues. Passing it through the logistic activation function then generates its expected production probability.

We then calculated the difference in activation between the frequent suffix and the infrequent suffix given the semantic cues and the stem cue present on a particular trial. This corresponds to the expected effect of frequency on the log odds of choosing these suffixes when presented with a particular meaning and a particular stem. For example, the activation *nem* receives when the subject is presented with its original meaning is the sum of activations from DIMINUTIVE, SINGULAR and the stem. It is the log odds of *nem* in its original meaning. If the frequent suffix is *dan*, then the activation of *nem* would be subtracted from the activation *dan* would receive in its original meaning, i.e., from BIG, PLURAL and the same stem. If *nem* is the frequent suffix, then the subtraction is in the opposite direction. For the novel meaning, the semantic cues are DIMINITUVE and PLURAL for both suffixes. This results in a single value for each trial that is the model's expected log odds of the frequent suffix relative to the infrequent suffix given the stem and meaning (original or novel) on that trial.

In addition to the simplest form of the connectionist model, we also evaluated more complex models that allowed for configural cues as well as different weighting of semantic vs phonetic cues. This is well-motivated by previous work. For example, there is evidence that linguistic units can have associations that their parts do not have (Houghton, 2025; Kapatsinski, 2009, 2023; see also Saavedra, 1975 for stimuli in classical conditioning). Additionally, there is also evidence that adult native English speakers show greater reliance on semantic cues than phonological cues in production (Culbertson et al., 2018). For these models, there was qualitatively a lot more variation across different parameter values. In order to address this, we conducted a grid search to determine the best values for each of the model parameters (the salience of the phonetic cues, the salience of the semantic cues, and the salience of outcomes). The ground truth of this grid search was the model estimates for the human data across condition, stem familiarity, and meaning. Model predictions were compared using mean squared error.

We evaluated the discrepancies between model predictions and human data by using expected log odds from the model as predictors in a Bayesian logistic regression models. In each model, the dependent variable was whether the human learner, for a given trial, selected the frequent suffix or the infrequent suffix.

Specifically, we used expected difference in log odds (referred to as freq_minus_infreq) with random intercepts for participant and stem identity (Equation 6). This model serves as a baseline. In our second model, we calculated separately the activation received by the suffix from the stem and from the meaning (using the perceptron). The motivation behind this is that the perceptron is agnostic to whether the cue is a stem or a meaning, but participants may show varying sensitivity to one over the other (e.g., Gagliardi et al., 2017 argue that learners are more sensitive to form cues than to semantic cues). Using the activation from the stem and the meaning as separate predictors allows the statistical model to fit different slopes for the stem activations and the meaning activations. We then added a fixed-effect for condition as well (Equation 7). Finally, our third model was identical to the second model except we included an interaction effect between condition and the stem activations, and an interaction between the meaning activations and condition (Equation 8). This model tests whether the perceptron captures the effects of type and token frequency and their interactions with meaning and stem novelty, or if some effects are not captured by the perceptron and therefore provide support for additional mechanisms.

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

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

$$\begin{aligned} \text{frequency} \sim\ & \text{freq\_minus\_infreq\_stem} + \text{freq\_minus\_infreq\_meaning} + \text{condition} \\ & + \text{freq\_minus\_infreq\_stem:condition} + \text{freq\_minus\_infreq\_meaning:condition} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{8}$$

Additionally, in order to examine the effect of configural cues and modified semantic salience, we compared five additional models: one with configural cues and modified semantic salience, one with configural cues and modified semantic salience, both of which were allowed to vary by condition, one with configural cues, increased semantic salience, and the original statistical predictors (condition:meaning:stem_condition), one with only configural cues (no modified semantic salience) and one with only modified semantic salience (no configural cues).

We then compared these with the model with only statistical predictors, and the three models described previously (Equation 6, Equation 7, and Equation 8).

We then compared these models using leave-one-out-cross-validation (Table 6). The results of leave-one-out cross-validation suggest that the models that include configural cues or increased sensitivity to semantic cues outperform all other models as well as the model with only statistical predictors. Further, allowing the configural cues and increased semantic saliency to vary by condition does not result in further improvement of the model. Among those models, any of the models that include special salience of semantic cues perform better than any model that does not. In order to visualize this, we also included a side-by-side plot of the human data with the RW activations from the model that included both modified semantic salience and configural cues (Figure 9). For comparison, we also plotted the activations from the model without modified semantic salience or configural cues.

Overall, the results of our simulations suggest that a simple connectionist model fits the human data well. Further, a model with configural cues and increased salience of semantic cues fits the human data better than even the original statistical predictors. This parallels previous findings in

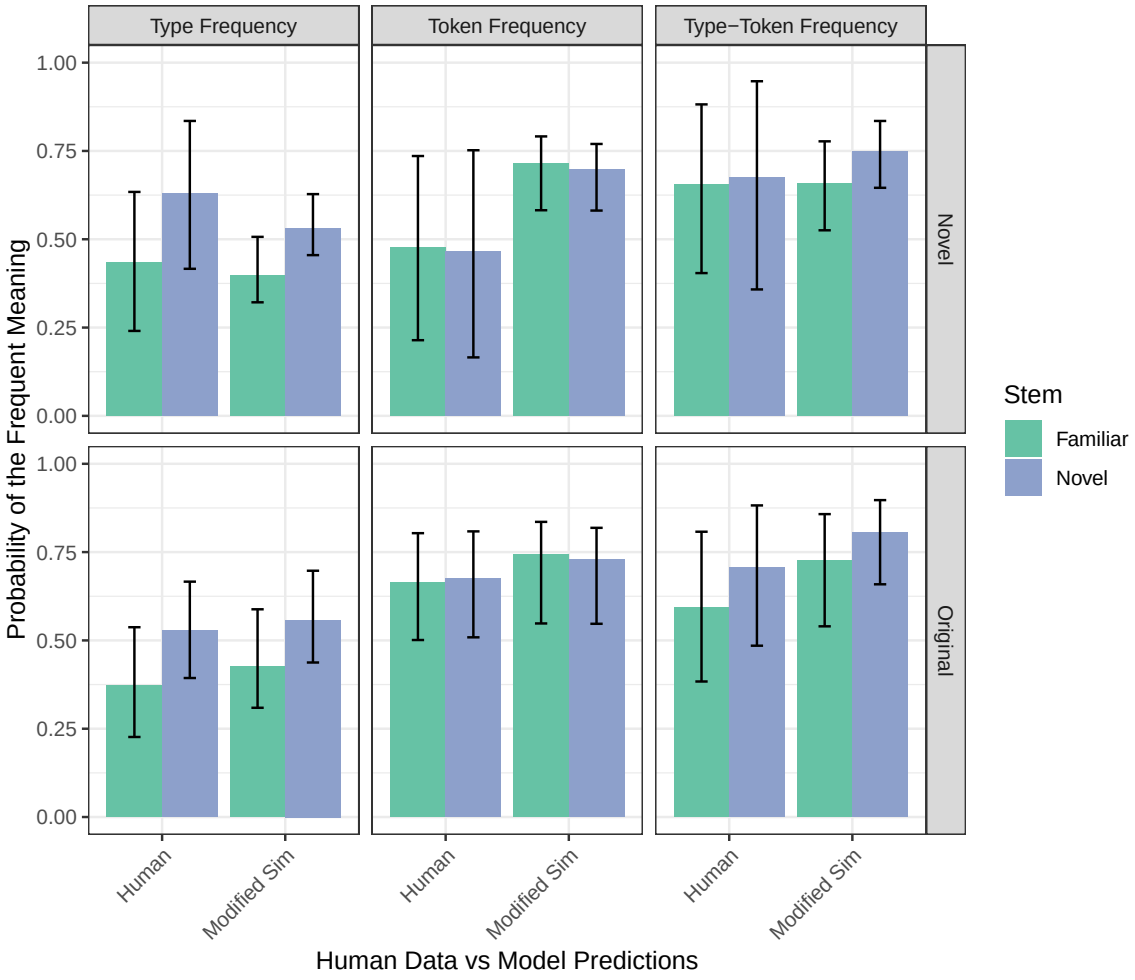

Figure 9: Plot of the original RW suffix activations and the modified RW activations (configural cues plus modified salience of semantic cues) versus the human results. The y-axis is the probability of producing the frequent meaning. The top facet is original meanings and the bottom axis is novel meanings. Along the x-axis, 'Modified Sim' corresponds to the RW simulations with configural cues and increased semantic saliency. 'RW Logistic' corresponds to our original suffix activations. 'Human' corresponds to the original human data. The visualization shows that the original RW suffix activations seem to fall short in predicting the human data for the token frequency condition.

Table 6: Results of the leave-one-out cross-validation for the additional simulations.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Configural Cues, Increased Saliency of Semantic Cues, and Original Statistical Predictors	0.00	0.00	-284.65	19.26
Configural Cues and Increased Saliency of Semantic Cues	0.00	1.71	-284.65	19.14
Configural Cues and Increased Saliency of Semantic Cues Varying by Condition	-0.19	2.40	-284.84	18.92
Only Increased Saliency of Semantic Cues	-4.01	2.13	-288.66	19.75
Only Configural Cues	-65.15	8.94	-349.80	23.49
Original Statistical Predictors	-1280.46	28.00	-1565.11	26.03
Stem and Meaning Activations by Condition	-1285.76	27.49	-1570.41	25.21
Only RW Activation	-1299.87	27.10	-1584.52	24.81
Stem and Meaning Activations (not by Condition)	-1300.03	27.14	-1584.68	24.83

demonstrating that humans learn associations for configural cues and that humans are more sensitive to semantic cues than phonological cues, and provides support for a connectionist approach to human cognition.

Comprehension Models

In this section, we present the results of modeling the comprehension data using the Rescorla-Wagner logistic model. Similar to the production data, the RW model was trained on the same series of training trials as our human subjects (presented in a random order). However, instead of modeling the data with the cues being stem identity and meaning, since we were modeling comprehension the cues instead were the steam identity and suffix. Additionally, the outcome was the meaning.

For each of the items in the comprehension task, we computed the activation of that item for the novel meaning (diminutive plural) and the familiar meaning (big plural for *dan* and diminutive singular for *nem*). We then calculated the difference in activation between the new meaning and the familiar meaning. This served as our value for the RW prediction. A larger value indicates the RW model predicts a higher probability of choosing the novel meaning while a smaller (more negative) value indicates the RW model predicts a higher probability of choosing the familiar meaning.

In order to examine the fit of the RW model to the human data, we ran six Bayesian logistic regression models. The dependent variable for all the models was whether the participant chose the novel meaning (1) or the familiar meaning (0). The independent variables varied across models. For the first model, the independent variable was simply the RW predictions (new_minus_old). In the second model, we also included condition (baseline, High Type, and Low Token) along with an interaction between condition and the RW predictions. In the third model, instead of condition we included stem familiarity (and interaction). The fourth model included all three (along with the two-way interactions and the three-way interaction). The final two models included only the original statistical predictors instead the RW predictors. More specifically, the fifth model included only condition, and the sixth model included condition and stem familiarity (along with their interaction). All six models included a random intercept for participant and a random intercept for stem identity. The model equations for each are included below.

$$\begin{aligned} \text{meaning} \sim\ & \text{new\_minus\_old} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{9}$$

$$\begin{aligned} \text{meaning} \sim\ & \text{new\_minus\_old} * \text{condition} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{10}$$

$$\begin{aligned} \text{meaning} \sim\ & \text{new\_minus\_old} * \text{stem\_familiarity} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{11}$$

$$\begin{aligned} \text{meaning} \sim\ & \text{new\_minus\_old} * \text{condition} * \text{stem\_familiarity} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{12}$$

$$\begin{aligned} \text{meaning} \sim\ & \text{condition} \\ & + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{13}$$

Table 7: Results of the leave-one-out cross-validation for the additional simulation. The results suggest that the RW predictions capture the effect of stem familiarity (because adding stem familiarity does not improve the model fit much) but the model is not fully capturing the effect of condition (because adding condition does improve the model fit).

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
RW Predictions, Condition, and Stem Familiarity	0.00	0.00	-745.02	25.24
RW Predictions and Condition	-5.55	3.44	-750.57	25.16
RW Predictions and Stem Familiarity	-71.66	10.92	-816.69	23.32
Only Condition	-74.29	10.93	-819.31	23.43
Condition and Stem Familiarity	-77.47	11.89	-822.50	23.73
Only RW Predictions	-85.58	12.46	-830.60	23.53

$$\begin{aligned} \text{meaning} \sim \text{condition} * \text{stem\_familiarity} \\ + (1|\text{participant}) + (1|\text{resp\_stem}) \end{aligned} \tag{14}$$

In order to examine which model fit the human data best, we once again turned to leave-one-out cross-validation. The results are presented below in Table 7. The results suggest that the RW predictions capture the effect of stem familiarity (because adding stem familiarity does not improve the model fit much) but the model is not fully capturing the effect of condition (because adding condition does improve the model fit).

In order to further understand where the model is falling short, we present the results of the model presented in Equation 11 (presented in Table 8 and visualized in Figure 10). The visualization demonstrates that the model predicts the Low Token condition (3-3) well, but falls short in its predictions for the other two conditions (baseline and High Type), demonstrated by the interaction effect between the low token condition and the RW predictions. Qualitatively, this appears to be due to the model predicting a strong effect of stem familiarity in the High Type condition and a minor effect of stem familiarity in the baseline condition (as demonstrated in Figure 10) while humans do not show any effect of stem familiarity.

Table 8: Results of the model in Equation 11 containing the predictions from the Rescorla-Wagner logistic model along with condition and stem familiarity. The results suggest the model does a good job of explaining the human data for the Low Token condition, but not for the High Type or baseline conditions.

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept (Baseline/12-3)	-2.84	0.37	-3.59	-2.12	0.00
High Type (12-12)	1.76	0.38	1.11	2.62	100.00
Low Token (3-3)	4.34	0.48	3.38	5.27	100.00
RW Predictions	0.48	1.12	-1.55	2.94	66.66
High Type:RW Predictions	1.66	1.61	-0.76	5.52	88.12
Low Token:RW Predictions	15.03	2.66	9.57	20.06	100.00

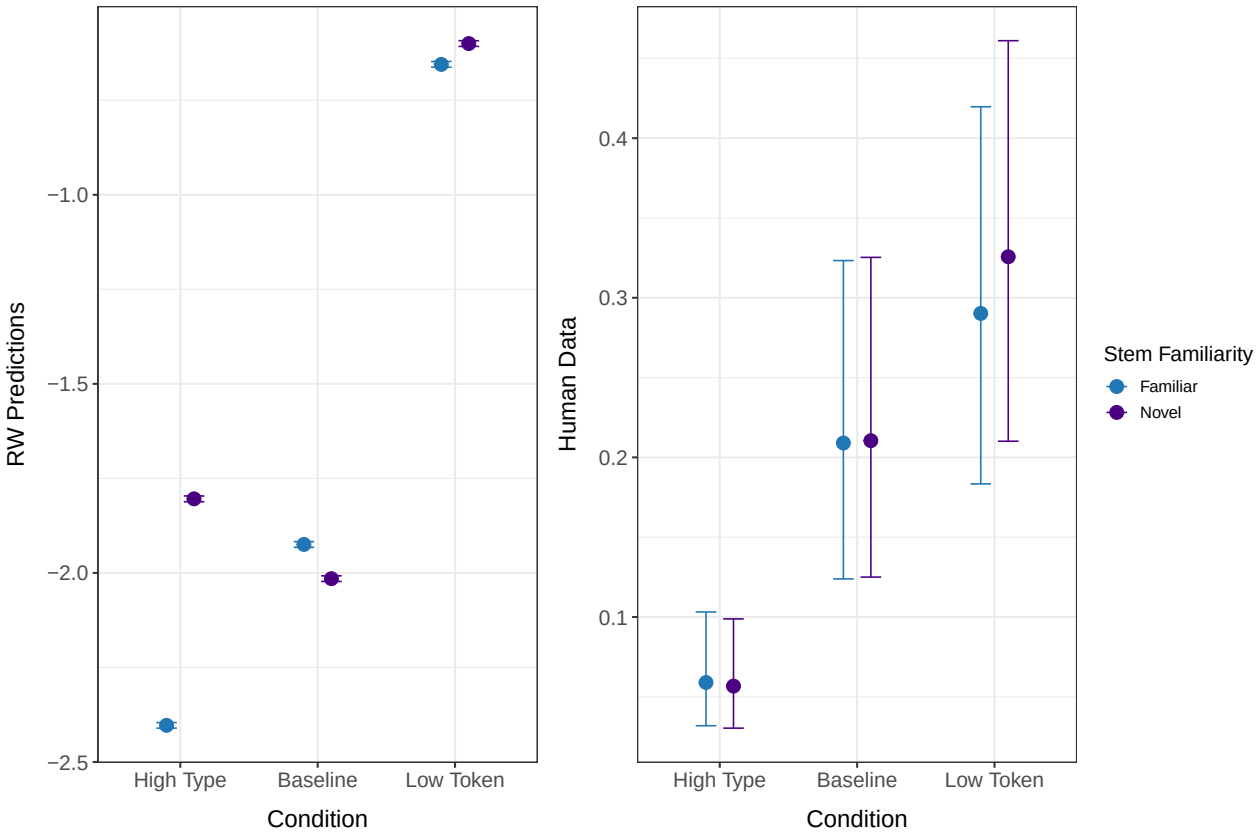

Figure 10: Plot of the RW predictions as a function of condition and stem familiarity (left) and human results as a function of condition and stem familiarity (right). The x-axis denotes the condition and the coloring denotes whether the stem was novel (purple) or familiar (blue).

Discussion

Our results suggest that in production, having a high type and high token frequency together lead to a preference for the frequent suffix, and this preference is even stronger if the stem is novel. On the other hand, when there is a high type frequency but not a high token frequency, there is only a preference for the frequent form when the stem and meaning are both novel. Finally, for token frequency there is only a preference for the frequent form when the meaning is original. Overall, these results suggest that type frequency and token frequency both play a role in semantic extension, with a high type frequency encouraging semantic extension while a high token frequency encourages using the suffix to communicate the original meaning. Similar to Harmon & Kapatsinski (2017), these effects are mitigated when both forms are brought into memory.

A similar pattern emerges in comprehension. When there is a high token-to-type ratio, the original meaning is preferred more. However, when there are as many types as there are tokens, there is an increased preference for the novel meaning relative to when there is a high token-to-type ratio.

Our results also demonstrate that the Rescorla-Wagner model with a logistic activation function is able to capture the human data. Further, the model matches the human data best when it includes configural cues and an increased salience for semantic cues relative to phonological cues.

Caballero, G., & Kapatsinski, V. (2022). How agglutinative? Searching for cues to meaning in choguita rarámuri (tarahumara) using discriminative learning. *Morphological Diversity and Linguistic Cognition*, 121160. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E6D16CD11DD783B5AA2B0DEE7C0CC285

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2018). *Do children privilege phonological cues in noun class learning? 40*. https://escholarship.org/uc/item/74g7g684

Dawson, M. R. (2008). *Minds and machines: Connectionism and psychological modeling*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=7ZHD6uOqwcsC&oi=fnd&pg=PR5&dq=Minds+and+machines:+Connectionism+and+psychological+modeling&ots=LClXwhWiWV&sig=_e5KSCaNIE5KDZ8bN3ilfzOxEI8

del Prado Martın, F. M., Ernestus, M., & Baayen, R. H. (2004). Do type and token effects reflect different mechanisms? Connectionist modeling of dutch past-tense formation and final devoicing. *Brain and Language*, *90*(1-3), 287298. https://www.sciencedirect.com/science/article/pii/S0093934X03004413?casa_token=X0BstCP27VoAAAAA:W1dhSKyCw5nw0yF6H9xcDiYFWnDDpwBEaX7TC_LT-U0

Gagliardi, A., Feldman, N. H., & Lidz, J. (2017). Modeling Statistical Insensitivity: Sources of Suboptimal Behavior. *Cognitive Science*, *41*(1), 188–217. https://doi.org/10.1111/cogs.12373

Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalisation in connectionist networks. *Language and Cognitive Processes*, *10*(6), 601–630. https://doi.org/10.1080/01690969508407115

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

Houghton, Z. N. (2025). *Multi-word representations in minds and models: Investigating the storage of multi-word phrases in humans and large language models* [PhD thesis]. https://search.proquest.com/openview/59e91c3252682d11c048247ba152f037/1?pq-origsite=gscholar&cbl=18750&diss=y

Kapatsinski, V. (2009). Testing theories of linguistic constituency with configural learning: The case of the english syllable. *Language*, *85*(2), 248–277. https://doi.org/10.1353/lan.0.0118

Kapatsinski, V. (2023). 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

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73193. https://www.sciencedirect.com/science/article/pii/0010027788900327

Rescorla, R. A., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning II, Current Research and Theory*, *Vol. 2*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533536. https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/323533a0&casa_token=VawPIGxM4JkAAAAA:ZljCRrcfNDV9T6R3B4bcMzU9jQ2HTRuN4MTa15ZDdMtRSr_5vrcbaSBO2VrGcfIpo4xJsxdcnSMX5XWUTQ

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. *Psycholinguistics: Critical Concepts in Psychology*, *4*, 216271. https://books.google.com/books?hl=en&lr=&id=n1cd70T4WxIC&oi=fnd&pg=PA221&dq=rumelhart+and+mcclelland+1986&ots=1_jE4qNRQr&sig=_XcVec9KMXJspJllbk9NJ9IYGug

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation*, *6*(3), 314326. https://www.sciencedirect.com/science/article/pii/0023969075900120

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245. https://psycnet.apa.org/record/2007-05396-002