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 the language (mainly type frequency and token frequency) affect which suffixes are extended to express the diminutive plural meaning.

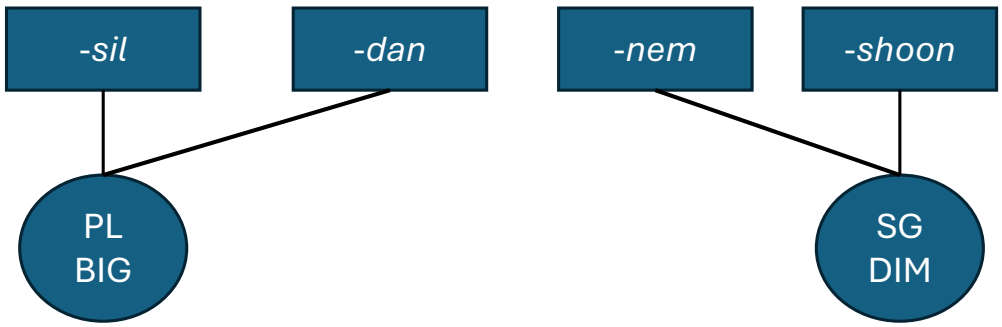

Figure 1: A description of the suffixes in our artificial languages. The thicker lines denote the more frequent form in each language: the plural *-dan*$_{PL}$ in the Dan language and the diminutive *-nem*$_{DIM}$ in the Nem language.

During the exposure phase, each suffix was 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

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

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 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 unaffixed 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

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 linear mixed-effects model on the production data. The depen-

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

dent variable was whether the participant produced the frequent suffix. The frequent suffix was coded as 1 if the meaning was diminutive plural and the suffix used was the frequent one (*dan* in the Dan language, *nem* in the Nem language). However, if the meaning was big plural or diminutive singular, the frequent suffix was coded as 1 if the suffix chosen was *dan* for big plural and 0 if the suffix chosen was *nem* for diminutive singular in the Dan language, and vice versa in the Nem language.

We treatment coded condition such that the intercept was the type-token frequency condition. Thus, a larger intercept indicates that the frequent suffix was more likely to be produced than the infrequent suffix when it had a high type and token frequency. A larger coefficient estimate 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 condition as sum-coded variables. Meaning had two values, either original or 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 condition similarly had two values, 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 condition 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.

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

We find a meaningful effect for the intercept, suggesting that the frequent suffix is chosen more than the infrequent suffix when there is a high type and high token frequency. We also find a meaningful effect for novel stems, suggesting that in general the frequent suffix is produced more than the infrequent suffix when the stem is novel. Perhaps most interestingly, we find a meaningful interaction between the type frequency condition and novel stem, suggesting that in the type frequency condition, there is a preference for the frequent suffix over the infrequent suffix when the stem is

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

novel but not when it is familiar. On the other hand, no such interaction effect is found in the token frequency condition. Further, we find a three-way interaction between type frequency condition, novel meaning, and novel stem, suggesting that learners are more 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. Additionally, when the frequent suffix has a higher type frequency than the infrequent suffix, learners are more likely to use it if the stem is novel, but not when the stem is familiar. Finally, in general there is no preference to use the frequent suffix more for novel meanings if the suffix has only a higher token frequency than the infrequent suffix.

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. Analogous to the first production model, the independent variables were condi-

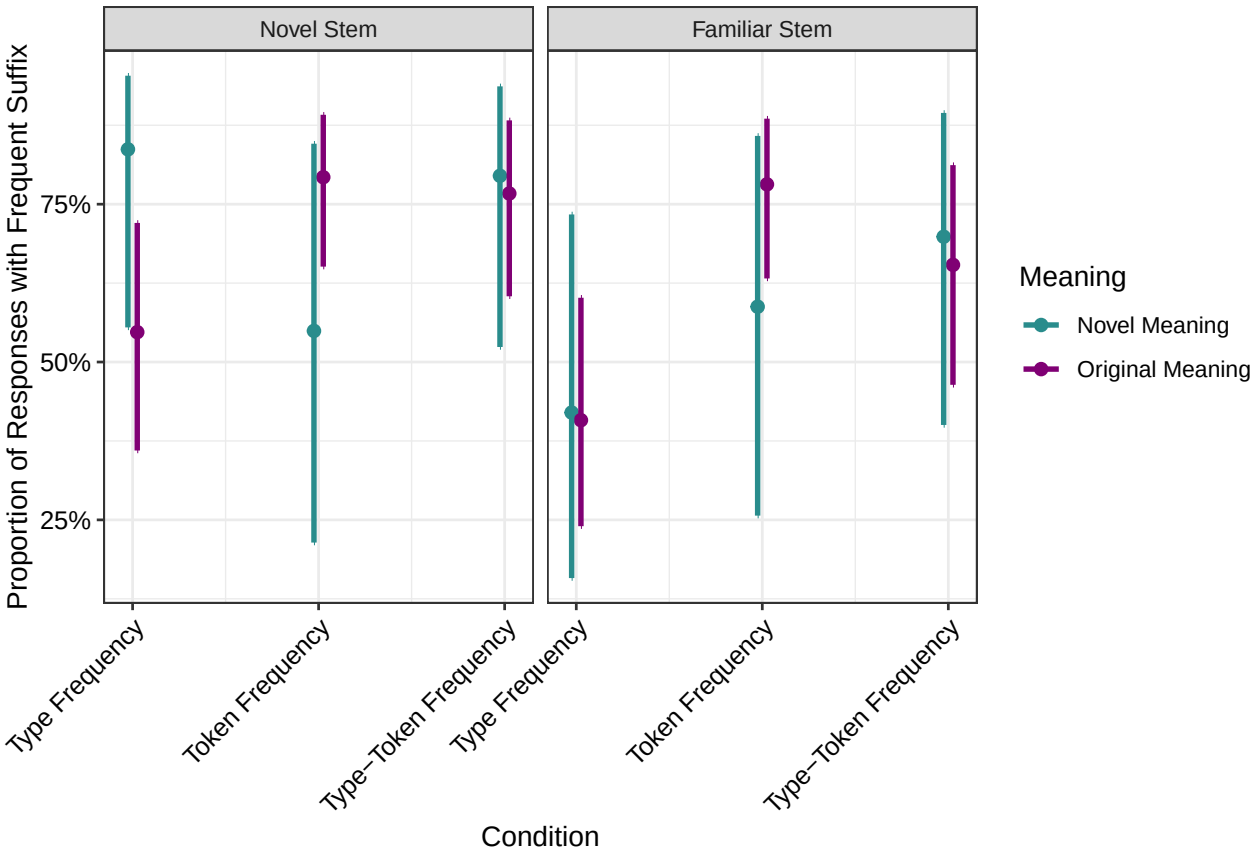

Figure 2: Plot of the statistical model estimates for the production task. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while purple points corresponds to the original meaning. The facet indicates whether the stem was novel or familiar.

Table 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

tion, meaning, and stem condition. Once again, condition was treatment coded such that type-token frequency was the reference level and meaning and stem condition were sum-coded. The random-effects structure was identical to that in (**eq_prodmodelstem?**). The model results are presented in Table 3 and visualized in Figure 3.

In general we find no meaningful effects except for a small 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 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).

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

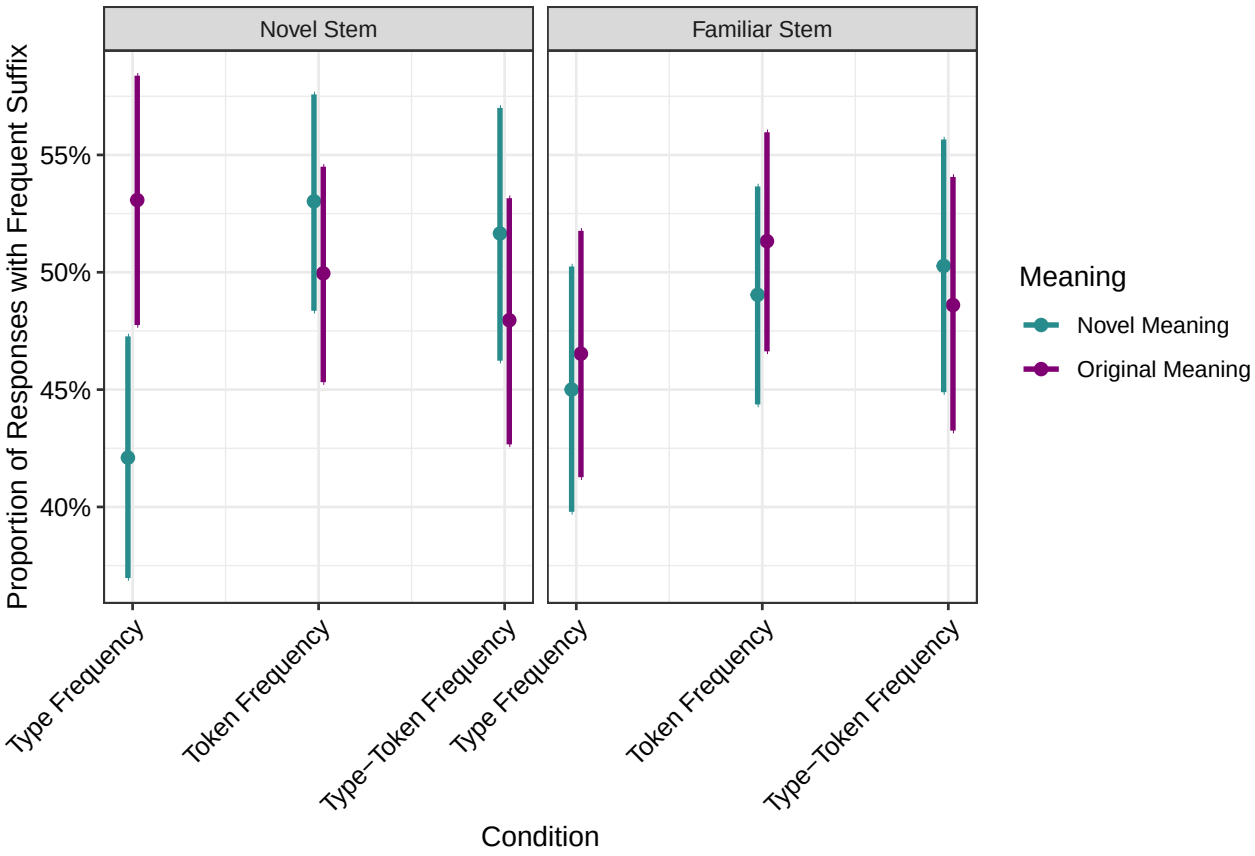

Figure 3: Plot of the statistical model estimates for the form-choice task. The x-axis indicates the condition (type frequency, token frequency, or type-token frequency). The y-axis corresponds to the proportion of responses with a frequent suffix. Blue points indicate that the meaning was novel while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

meaning while a negative estimate indicates that participants were more likely to choose the original meaning. Unlike the other tasks, our independent variable was coded quite differently. Specifically, our conditions varied in what the token-to-type frequency was. 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. Thus, instead of dividing by condition, we included the token-to-type ratio as a fixed-effect. Since this approach maximizes 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. 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.

We treatment coded condition (referred to as condition_numeric) such that the intercept was the 12-3 condition. As a result, 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. Our results for the comprehension task are included in Table 4 and visualized in Figure 4.

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

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 than when there is a higher token-to-type ratio. 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.

Table 4: 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

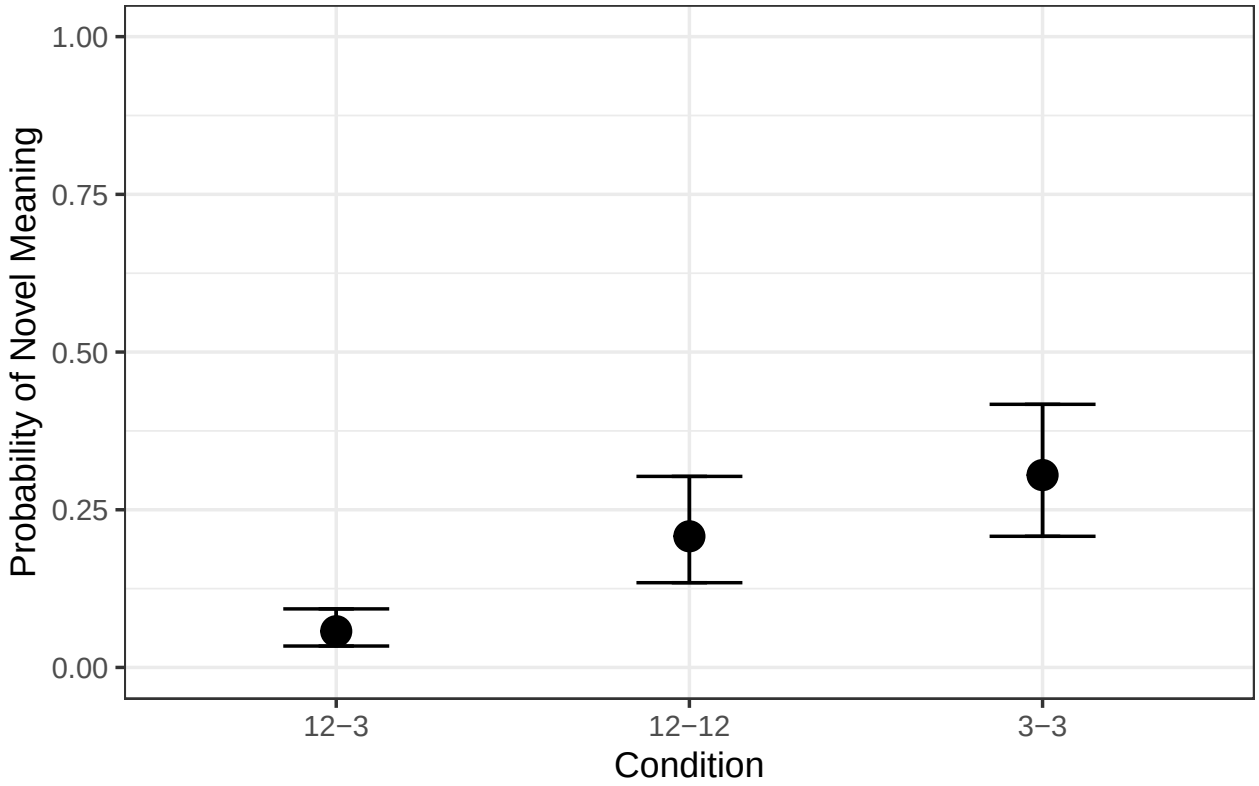

Figure 4: Plot of the model estimates for the comprehension task.

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

Similar to the form-choice task, we also ran a model with stem familiarity as a fixed-effect. The dependent variable remained the same (whether participants chose the original or novel meaning), however in addition to condition, we also included stem condition (familiar or novel) as a fixed-effect, along with its interaction with condition. A random intercept for participant was also included.

The results of the model are presented in Table 5. Our results show no effect of stem familiarity on comprehension.

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 type-to-token ratio decreases the preference for the original meaning, it does not eliminate it.

Rescorla-Wagner Model

In this section we examine whether the Rescorla-Wagner model (Rescorla & Wagner, 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, 2023; Rescorla & Wagner, 1972). 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, 2023). λ_{max} is the learning rate.

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

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

While the Rescorla-Wagner model is often used with a linear activation function, in the present simulations we use a logistic activation instead (Equation 5) . The motivation behind this decision is that the Rescorla-Wagner model with a logistic activation function is more sensitive to type frequency than the Rescorla-Wagner model with a linear activation function (Caballero & Kapatsinski, 2022). Further, as Kapatsinski (2023) demonstrates, using a logistic activation function mitigates spurious excitement, a side-effect of using the linear activation function whereby an absent cue and an absent outcome become associated with eachother.

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

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, λ_{min} was set to 0, and λ_{max} 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. This was done by feeding the model each trial of the exposure phase in a random order. The cues were the 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*). The model then learned associations between each cue and each outcome.

Next, for each of the items in the production task, we summed the associations for each cue-outcome present in that trial. For example, given the item *baldan* and the meaning big plural, we consulted the matrix of cue-outcome associations for *bal* and *dan*, big and *dan*, and plural and *dan*. We then summed these to get the predicted associations for *bal* with the big.pl meaning and *dan*. This resulted in a matrix that included the model's learned association strength for any given cue with each of the four suffixes.

Following this, in order to compare the simulations with the human data, we calculated association strengths depending on whether the meaning was original or familiar. For the original meanings, we subtracted the activation strengths between the original meaning and the appropriate suffix. Specifically, this is the difference in activation strength between the meaning cue big plural and the outcome *dan*, and the meaning cue diminutive singular and the outcome *nem*. If the language was dan (i.e., *dan* was the frequent suffix), then the activation strength between diminutive singular and *nem* was subtracted from the activation strength between big plural and *dan*. If the language was nem (i.e., *nem* was the frequent suffix), then the subtraction was in the opposite direction. To calculate the association strengths for novel meanings, if the frequent suffix was *dan* then the activation strength between diminutive plural and *nem* was subtracted from the activation strength between diminutive plural and *dan*. If the frequent suffix was *nem* then the opposite was done. This results in a single value

for each trial that is the model's predicted activation of the frequent suffix relative to the infrequent suffix.

In order to test our model, we evaluated whether the model is a good predictor of the human data. Thus, we ran two 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. The first model was a simple model that modeled the human data as a function of the Rescorla-Wagner activation strength (referred to as `freq_minus_infreq` in the model syntax), with random intercepts for participant and stem identity (Equation 6). In our second model, we calculated separately the stem activation and the meaning activation (using the RW model). The motivation behind this is that the Rescorla-Wagner model is agnostic to whether the cue is a stem or a meaning (because it originally has no way of knowing whether a given cue is part of the stem or the meaning), but participants may show varying sensitivity to one over the other. Modeling it in this manner allows the statistical model to fit different slopes for the stem activations and the meaning activations. We also included fixed-effect for condition as well as an interaction effect between condition and the stem activations, and an interaction between the meaning activations and condition (Equation 7).

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

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

The results of the models are presented in Table 6 and Table 7.

Overall we find that the general RW activations for both stem and meaning are able to account for the human data. Further, we find that when token frequency is high, meaning activations are

Table 6: Results of the statistical model with only the RW predictions (Equation 6).

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.34	0.31	-0.28	0.96
RW Predictions	0.38	0.09	0.20	0.56

Table 7: Results of the statistical analysis containing the separated stem and meaning activations and condition as a fixed-effect (eq-rwmodel3).

	Estimate	Est.Error	Q2.5	Q97.5	% Samples > 0
Intercept	0.38	0.62	-0.83	1.60	73.10
Stem Activations	0.60	0.45	-0.26	1.51	91.60
Meaning Activations	0.33	0.48	-0.61	1.28	75.04
Type Frequency	-0.28	0.66	-1.61	0.99	33.54
Token Frequency	-1.80	0.92	-3.72	-0.16	1.45
Stem Activations:Type Frequency	0.01	0.46	-0.92	0.89	51.56
Stem Activations:Token Frequency	-0.66	0.47	-1.61	0.25	7.88
Meaning Activations:Type Frequency	-0.40	1.50	-3.79	2.22	40.20
Meaning Activations:Token Frequency	2.75	0.77	1.30	4.29	99.99

a better predictor (evident by the interaction effect between Meaning Activations and Token Frequency in Table 7).

In order to compare which model best fits 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 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.

We performed leave-one-out cross validation for each of the two models with the RW pre-

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

Table 8: Results of our leave-one-out cross-validation.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Original Statistical Predictors	-0.26	0.77	-1564.93	26.05
Stem and Meaning Activations by Condition	-2.80	3.96	-1567.47	25.58
Only RW Activation	-18.47	6.41	-1583.14	24.93

dictors as well as the model with just the human predictors. The results of our leave-one-out cross-validation are included in Table 8. When using leave-one-out cross-validation, a model can be considered as outperforming another model if the difference in elpd is greater than the standard error of the elpd. In the case of our models, because all of the elpd differences are less than the standard error of the elpd, it is difficult to draw conclusions about which model fits the data best.

Additional Simulations

There is evidence that humans learn associations for configural cues that their parts do not have (Kapatsinski, 2009). For example, in English syllables, rimes are treated as a single constituent (Figure 5) and Kapatsinski (2009) demonstrated that native English speakers can learn to associate rimes with an outcome even in cases when the onset and nucleus are associated with different outcomes.

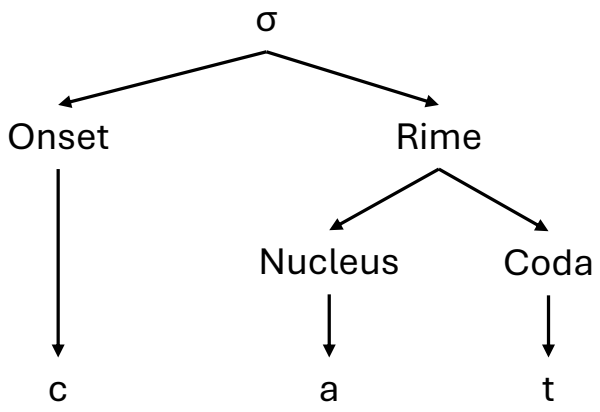

Figure 5: A visualization of English syllable structure.

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 example, Culbertson et

al. (2018) demonstrated that in learning an artificial noun class adults rely more on semantic cues than phonological cues (whereas interestingly children rely more on phonological cues than semantic cues).

Thus, in order to further examine the differences between the human and model predictions we made two modifications to the simulations. First, instead of modeling the suffix meanings (diminutive plural, diminutive singular, big plural, big singular) as two separate cues, we also included a configural cue (analogous to an interaction effect in linear regression) that could be associated with the outcome. This configural cue was simply a combination of the two meanings. For example, given the meaning "big plural", the cues comprised "big", "plural", and "big.plural". Similar to the previous simulations, the model was presented with these cues paired with one of the four suffixes. The model was trained on the same data that the humans were trained on.

Second, in addition to configural cues, we also increased the alpha value for semantic cues relative to the phonological cues. This results in semantic cues being more associable with outcomes than phonological cues. Specifically, alpha was 0.05 for phonological cues and 0.25 for semantic cues. All other variables remained unchanged.

Analogous to the previous simulations, we then calculated the model's predicted activation for the frequent suffix. If the meaning was novel, this was simply the difference between the activation strength for the frequent suffix and the activation strength for the infrequent suffix. If the meaning was original, then it was the difference between activation strength of big plural with *dan* and the activation strength of diminutive singular and *nem* if *dan* was the frequent suffix and vice versa if *nem* was the frequent suffix.

In order to evaluate this simulation, we ran a Bayesian logistic regression model with the human response as the dependent variable (1 if they chose the frequent suffix and 0 if they chose the infrequent suffix) and the model's prediction as the independent variable with random intercepts for participant and stem. We ran four models, one with configural cues and modified semantic salience, one with configural cues, increased semantic salience, and the original statistical predictors ($condition *$

Table 9: Results of the leave-one-out cross-validation for the additional simulation. 'Modified RW' corresponds to the model with configural cues and increased saliency for semantic cues. 'RW Model' refers to the RW logistic model. 'Statistical Predictors' refers to the original statistical analysis that contains no RW model predictions.

	elpd_diff	se_diff	elpd_loo	se_elpd_loo
Configural Cues and Increased Saliency of Semantic Cues	0.00	0.00	-285.61	19.26
Configural Cues, Increased Saliency of Semantic Cues, and Original Statistical Predictors	-0.10	1.92	-285.71	19.39
Only Increased Saliency of Semantic Cues	-7.03	1.88	-292.64	20.00
Only Configural Cues	-66.86	9.89	-352.48	23.73
Original Statistical Predictors	-1279.44	28.05	-1565.05	26.02
Stem and Meaning Activations by Condition	-1281.86	27.85	-1567.47	25.58
Only RW Activation	-1297.53	27.22	-1583.14	24.93

$meaning * stem_condition$), one with only configural cues (no modified semantic saleince) and one with only modified semantic salience (no configural cues). We then compared these with the model with only statistical predictors, and the two models from the previous simulations (**?@eq-rwsim1** and **?@eq-rwsim3**).

We then compared these models using leave-one-out-cross-validation (Table 9).

The results of leave-one-out cross-validation suggest that the models that include configural cues or increased sensitivity to semantic cues outperform all the original models as well as the model with only statistical predictors. Among those four, any of the models that include semantic predictors perform better than the model that includes only configural cues. 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 6). 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 an error-driven associative learning model, specifically the Rescorla-Wagner model, fits the human data well. Further, a model with configural cues and an increased salience of semantic cues fits the human data better than even the original statistical predictors. This parallels previous findings in demonstrating that humans

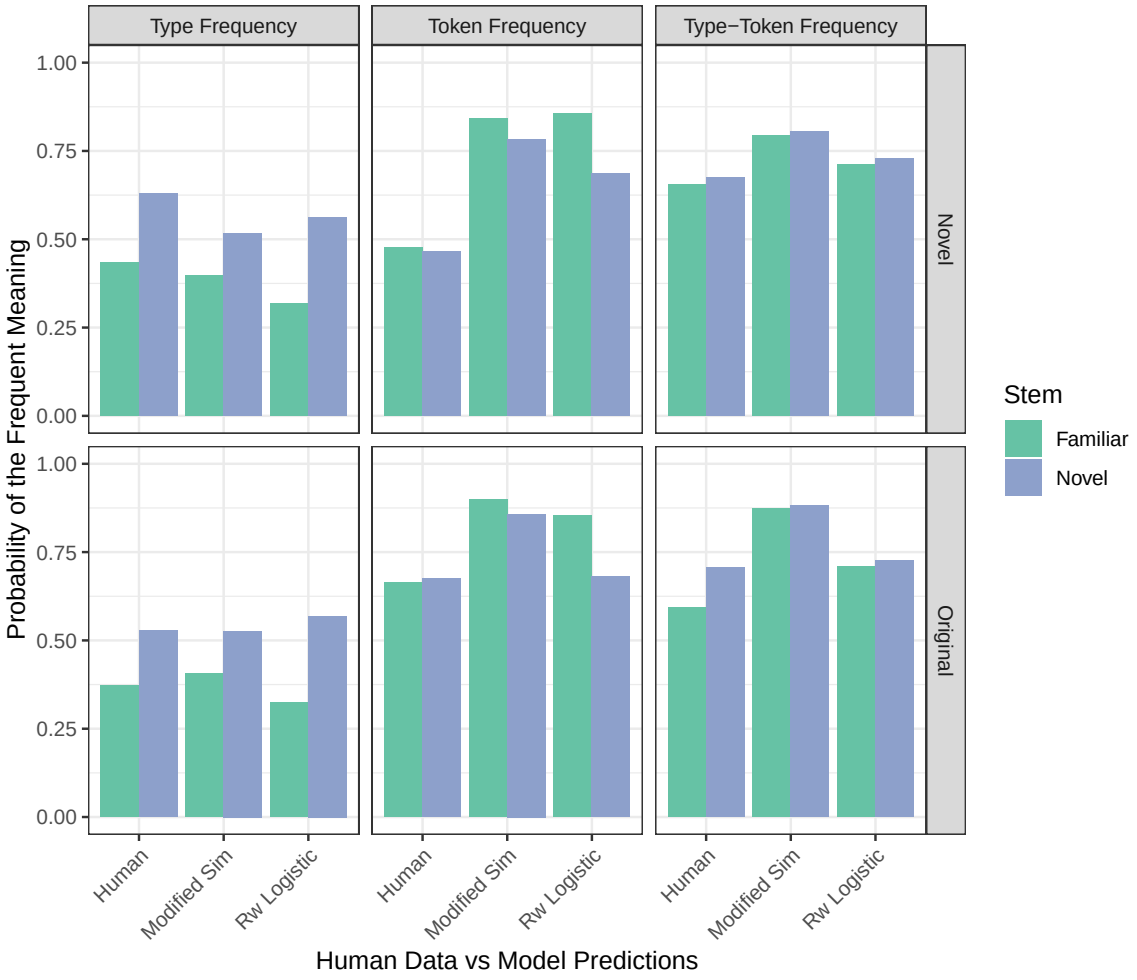

Figure 6: Plot of the original RW suffix activations and the modified RW activations (configural cues plus modified salience of semantic cues) versus the human results. The y-axis is the probability of producing the frequent meaning. The top facet is original meanings and the bottom axis is novel meanings. 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.

learn associations for configural cues and that humans are more sensitive to semantic cues than phonological cues.

Discussion

Our results suggest that in production,

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.

On the other hand, a similar pattern emerges. 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.

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

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

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. (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

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., & 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*.

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