Effects of Frequency and Distribution on Learning Minority Defaults

Katya Pertsova[1], Brandon Prickett[2], & Esther Chen[1]

[1]*University of North Carolina at Chapel Hill,* [2]*University of Massachusetts at Amherst*

1 Introduction

Many linguistic patterns can be described as having an “elsewhere” or default distribution. How such patterns are learned, particularly when it comes to minority defaults, has been a matter of debate. On some accounts, the existence of minority defaults is taken as evidence for symbolic rules (Pinker and Prince 1988, Marcus et. al. 1995). Others have argued that certain examples of minority defaults can be successfully handled by connectionist or associationist learning models that do not presuppose the existence of rules (Hare et al. 1995, Hahn and Nakisa 2000).

The two most discussed examples of minority defaults are German plurals and Arabic plurals. Both cases involve a plural allomorph which is infrequent compared to the other allomorphs, but which appears to have the default status: that is, it is extended to novel situations and words. For example, Marcus et. al. (1995) argue that the plural allomorph -s in German, while applying to a very small number of nouns in the language, is nevertheless extended to unassimilated borrowings, proper names homophonous with irregular nouns (e.g., Manns, *Männer), quoted nouns, titles, acronyms, truncations, nouns that are phonologically aberrant, and several other cases which are typically associated with regular inflections hypothesized to apply by default. However, the default status of -s has been questioned in several studies: while German speakers tend to generalize -s more frequently to non-rhymes (words that do not rhyme with any native word) compared to rhymes, the same can be said about another German plural suffix -(e)n (Zaretsky et al., 2016; McCurdy et al., 2020). Additionally, it is not the case that -s is the most commonly chosen allomorph for non-rhymes overall. Likewise, children do not seem to favor -s in the acquisition process. For instance, a longitudinal study of eight children by Bittner & Köpcke (2001) found that among the most common pluralization mistakes are lack of plural marking, followed by overgeneralization of -(e)n, followed by umlaut + -e. Other earlier studies report similar findings, with most common overgeneralization errors involving -(e)n, -e, umlaut, no ending, and only sometimes -s (Gawlitzek-Maiwald 1994, Park 1978, MacWhinney 1978). Some controversy surrounds the case of Arabic minority plural pattern as well (see Boudelaa & Gaskell 2002, Alshboul et al. 2012), but we will not get into the details of this pattern here.

Because there is no universally accepted example of a minority default in natural language, we turn to artificial languages to study how such patterns are learned and to examine the relative roles of type frequency and phonological distribution in the emergence of defaults. Artificial language learning (ALL) experiments provide an additional advantage of tightly controlling the learning environment and they have been used to investigate a similar question before (Nevat et. al. 2018). However, the usual caveat is that such experiments may not be directly informative about L1 acquisition (although they may resemble the early stages of L2 acquisition). To get a better understanding of how such patterns are learned, we also test an encoder-decoder RNN model (ED) and a rule-based model (Minimum Generalization Learner of Albright and Hayes, 2003) on the same artificial language and compare the results to human learning.

Results showed that participants successfully generalized the default pattern, even when it was a minority suffix. Explicit learners, who verbalized rules, were more likely to rely on phonological distribution, while implicit learners tended to overgeneralize the most frequent suffix. The Minimum Generalization Learner effectively captured minority defaults by representing them as a disjunction of specific rules, while the ED model struggled in the equal-frequency condition, highlighting potential limitations in its ability to abstract default patterns.

Proceedings of *AMP 2023/2024*

2 Previous Work

Several experimental and computational studies have investigated the question of learning minority defaults and the role of frequency vs. distribution in learning. Focusing on experimental work, studies on category learning in the domain of visual perception report that more diverse categories with a wide distribution are learned worse than those with a narrow, clustered distribution (Flannagan et al. 1986; Hahn et al. 2005), suggesting that defaults in general (patterns with a wide heterogeneous distribution) are at a disadvantage compared to natural-class patterns. A linguistic study by Nevat et al. (2018) examined the effects of type-frequency, morpho-phonological distribution, and predictability in an ALL experiment in which participants learned five plural suffixes, three of which had low frequency. The experiment took several days and a total of 6 training sessions. Initially, participants tended to overgeneralize the suffix with the highest frequency. Over time, however, they became more sensitive to the predictability of cues and began relying on phonological distribution. During the sixth session, type frequency and phonological distribution affected the participants' generalizations equally, and minority patterns were generalized as frequently as the majority ones. However, in this study low frequency was always correlated with wide distribution, and thus the effects of distribution alone were harder to interpret.

Computational work on minority defaults has found mixed results. While these kinds of patterns have often been used as evidence against connectionist theories of grammar (Pinker and Prince 1988, Marcus et. al. 1995), non-symbolic approaches have found some success in modeling them (Hare et. al. 1995, Hahn and Nakisa 2000). Here we look at results from two models: the rule-based Minimum Generalization Learner (Albright and Hayes, 2002) and a more recently-proposed neural network architecture: an encoder-decoder (Sutskever et al. 2014). Despite lacking any kind of symbolic component, the encoder-decoder neural network has shown success at modeling human behavior in a variety of tasks, such as English past-tense acquisition (Kirov and Cotterell 2018; see Corkery et al. for some caveats) and reduplication (Prickett et al. 2022). However, McCurdy et al. (2020) found that this kind of neural network failed to capture some aspects of human-like generalization (in particular generalization to unusual non-rhymes) when trained on German pluralization and attributed its failure to an inability to capture minority defaults.

3 Artificial Grammar Experiment

This artificial language learning study involved learning the distribution of three plural suffix-allomorphs conditioned by the phonological properties of the stem. Participants were first trained and then tested on using specific plural suffixes for nonce nouns. Two of the suffixes had a narrow, phonologically restricted distribution and one had an "elsewhere," wide distribution. In one condition all three suffixes had equal frequency, while in the other condition, the "elsewhere" suffix had the lowest frequency. Our main hypothesis was that both frequency and distribution will play a role in learning the elsewhere pattern, more specifically:

1. When frequencies of suffix-allomorphs are equal, the suffix with the elsewhere distribution:
    a. will be more often generalized to novel instances (especially those that don't match the narrow categories)
    b. will be learned slower and/or with less accuracy compared to suffixes with a narrow distribution
2. In the minority default condition, the most frequent narrow suffix:
    a. will interfere with learning the default and may be more often generalized to novel instances
    b. will be learned faster and with greater accuracy compared to other suffixes

3.1 *Stimuli* An artificial word-generator[1], modeled on English, was used to create three categories of nonce words, each defined by specific phonological properties conditioning plural suffix allomorphy. The three categories were:

[1] From this GitHub page: https://gist.github.com/wthe22/b963b0e6b3e581d073b2ad875f368f60

1. disyllabic words ending in a nasal consonant /m/ or /n/ (narrow category 1)
2. monosyllabic words ending in one of three clusters /nt/, /st/, /ft/ (narrow category 2)
3. words of 1, 2, or 3 syllables that do not end in nasal consonants or in -Ct clusters (elsewhere category)

Table 1. Example stimuli for three suffix categories

	Nasal Category	**Ct Category**	**Elsewhere**
Syllable count	2	1	Other
Word ending	-/n/ or -/m/	-/nt/, -/st/,-/ft/	Other
Example	*ranom* *pashem* *cotin*	*boft* *frest* *lunt*	*trofa* *nasp* *sopis*

Table 1 provides examples of stimuli for each of these three categories. The three suffixes assigned to these categories were -[jo], -[wa], and -[ler]. The audio-stimuli were generated by converting ARPABET transcriptions into *.wav* files using a synthesizer based on Google's Tacotron software modified by A. Aji (Wang et al., 2017).

Words that were three syllables long appeared only during testing. We assumed that since no three-syllable words were present in the training, these words would allow us to test how speakers generalize to words of a novel type. Overall, the testing data included words of three types: those that were seen in the training, novel 1 or 2-syllable words that followed the pattern of words in the training, and novel 3-syllable words which we assumed would be treated similarly to "non-rhymes" in the German studies, and which we predicted should take the default suffix.

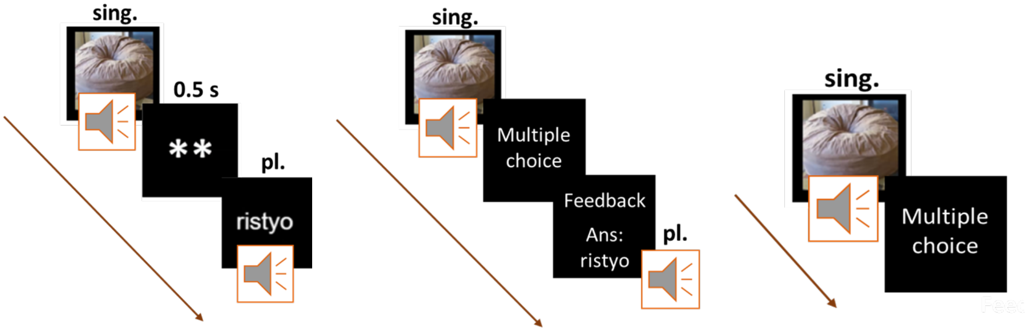
3.2 *Conditions* Participants were assigned to one of the two conditions: an *equal frequency* condition in which three categories were of equal frequency and a *minority default* condition in which the elsewhere category had the lowest frequency (see Table 2 for specific frequencies we used).

Table 2. Number of words for each category and condition

	Equal Frequency condition (39 words)	**Minority Default condition (40 words)**
Category 1	13 words (33.33%) narrow	18 words (45%) frequent/narrow
Category 2	13 words (33.33%) narrow	14 words (35%) mid-freq./narrow
Elsewhere Category	13 words (33.33%) default/wide	8 words (20%) default/wide

Categories 1 and 2 refer to the Nasal and -Ct categories in Table 1. In the minority default condition, there were two versions: one in which the Nasal category was most frequent, and one in which the -Ct category was most frequent. Additionally, within each condition, we varied which suffix was attached to which category, controlling for the possible effects of phonetic salience that a particular suffix may have.

3.3 *Procedure* Participants were told that they would be learning words in a made-up language. The experiment consisted of 2 training phases and a testing phase. Before each phase, participants were given a walk-through of what each trial would look like in the upcoming phase. Figure 1 provides a visual of the three phases and the respective trials. In Phase 1, participants heard a singular form (bare stem) of a nonce noun (e.g., *rist*) and saw a picture associated with it. Next, 2 asterisks appeared on the screen, followed by the plural form of the noun (e.g, *ristyo*), presented both auditorily and visually, written out in Latin characters. After hearing all training words once, participants advanced to Phase 2, which presented the same words but in a forced-choice task: upon hearing the singular form, participants had to choose the correct plural form given three possible options (corresponding to the three suffixes) and they received correct/incorrect feedback. The testing phase was like Phase 2 but without the corrective feedback.



Training Phase 1 Training Phase 2 Testing phase

Figure 1. Experimental procedure

After the study, participants filled out a questionnaire asking them to describe any rules or patterns they discovered during the experiment. The goal of this questionnaire was to distinguish between participants who formulated explicit rules or engaged in hypothesis testing vs. those who relied mainly on their intuition. So, the questionnaire responses were analyzed to sort participants into two groups: those that stated a correct or partially correct rule accounting for at least one category ("rule-staters") and the rest ("non-staters"). Previous work found that the rule-staters roughly correspond to explicit learners who often show a different pattern of behavior compared to other participants (Moreton and Pertsova, 2016, 2023). This distinction matters because it could be that explicit and implicit learners have different biases when it comes to learning specific patterns. For example, explicit learners may be particularly adept at learning the default distributions. Thus, we planned to examine whether rule-staters and non-staters would show specific differences in this experiment as well.

3.4 *Participants* The participants were recruited on the online platform for research-studies, *Prolific*. They were required to be native speakers of English, use English as their dominant language, and have no known visual or hearing deficits. A total of 101 participants took part in the study, but the results from 11 participants were excluded because they did not finish the study. Thirty of the participants were assigned to the Equal Frequency condition, and 60 were assigned to the Minority Default condition (30 to each sub-condition based on which feature was most frequent).

3.5 *Findings* The results were analyzed with respect to the proportion of correct responses for each suffix. Responses were counted as correct if they followed the following classification: words that ended in a nasal consonant were judged to belong to the nasal category, words that ended in -t were judged to belong to the -Ct category, and all other words were judged to belong to the Elsewhere category. This definition of correctness does not account for syllable count or word-final consonant clusters in our stimuli. We adopted such a method after realizing that almost all participants focused solely on the stem-final segment to make their suffix-choices. This was apparent from the rules they stated in the questionnaire and the fact that they treated 3-syllable words ending in a nasal or a -t as belonging to Category 1 and Category 2 (respectively). Since syllable count was fully correlated with the final segment for the narrow categories, and the -Ct cluster corresponded to the final -t, participants' responses aligned with the simplest hypothesis consistent with the data, so we evaluated correctness relative to this approach.

Analyzing participants' performance on the novel stimuli in the testing phase, we found that in both conditions, participants' suffix choices were guided by the phonological shape of the word. They chose the correct suffix most of the time for all three categories (see Figure 2). This supports hypothesis 1(a) that the default suffix will be correctly generalized to novel instances in the equal frequency condition. In the minority default-condition, we expected that instead the most frequent suffix would be overgeneralized to novel instances (hypothesis 2(a)). However, we found that the default was still learned well in this condition. A caveat must be noted: our original plan was to test whether participants would generalize the default to stimuli that were markedly different from those presented during training—namely, three-syllable words. However, as noted earlier, participants ignored syllable length differences. As a result, our study did not test whether participants could generalize the default to entirely novel or unusual forms. Instead, we only measured their generalization to novel words that closely resembled those seen in training.

We used multinomial logistic regression to model the probability of a correct response in both conditions with two main predictors: word-category (Nasal Category, -Ct category, Elsewhere) and distribution (wide vs. narrow categories). Other covariates were included in the model but were not significant predictors. The subject was included as a random effect. In both conditions, at least one of the narrow categories was learned more successfully than the wide Elsewhere category. More specifically, in the equal frequency condition, the Nasal category was learned significantly better than the Elsewhere category (odds=1.44, $p < 0.05$). And in the minority default condition, both the most frequent category and the second most frequent category were learned better than the Elsewhere category (odds=3.52, $p < 0.001$ and odds=2.34, $p < 0.001$, respectively). Additionally, the most frequent category was also learned significantly better than the second frequent category (odds=0.67, $p < 0.01$), confirming hypothesis 2(b) that the best learning will be observed with the most frequent suffix.

A Rao-Scott Chi-Square Test was used to compare learning of the default suffix across the two conditions. The difference turned out not to be significant ($X^2 = 0.072$ (1, $N = 1980$), $p = 0.93$). Likewise, there was no significant difference in how well Nasal category was learned in the two conditions ($X^2 = 1.2$ (1, $N = 1710$), $p = 0.29$). However, participants' performance on the -Ct category was significantly better in the equal frequency condition ($X^2 = 4.2$ (1, $N = 1710$), $p < 0.05$).

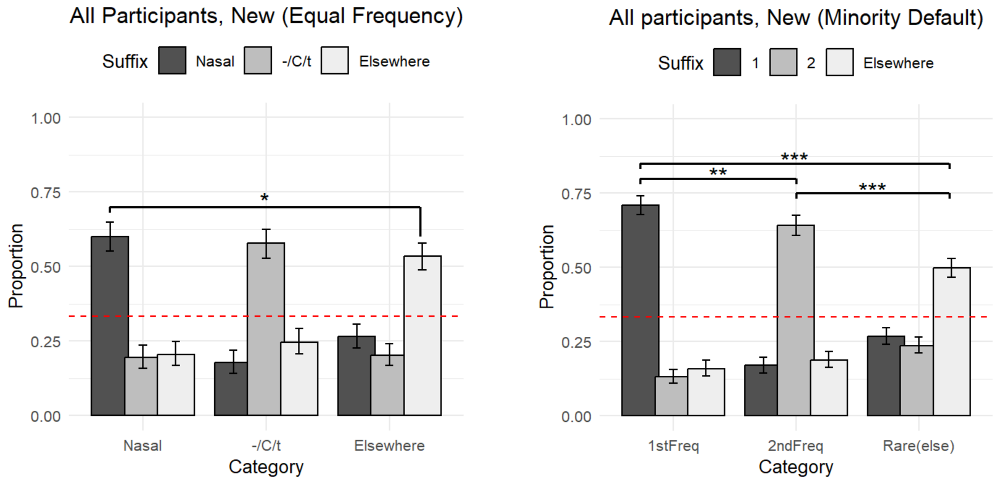


Figure 2: y-axix: Proportion of suffixes chosen by category for novel words. x-axis: Phonological category of a test word (e.g., ended in a nasal consonant, a -Ct cluster, or neither of those two) things). Dotted line shows

Thus, our overall findings show that participants were able to learn the distribution of the suffixes in both conditions and extend them to novel words, including the minority default condition. However, we were not able to test whether the learned default suffixes would apply "across the board" to any unusual example (as some rule-based models predict), because the stimuli we chose to test this prediction were not "unusual" enough for the speakers (i.e, the difference in word-length was ignored). In terms of the effects of frequency, as we predicted the more frequent suffixes were learned better than the less frequent minority suffix (in the minority default condition), but frequency did not completely trump the distribution in the sense that when faced with a word that did not fit any narrow categories, participants most often chose the default suffix, rather than the most frequent suffix. Additionally, when frequency was held constant, the wider default suffix was learned almost as well as the narrow suffixes, but slightly worse than one of the narrow suffixes.

3.5.1 *Explicit vs. Implicit Learners* In this section we consider whether the findings of the experiment change if we separate participants based on whether they were able to verbalize any rules after the experiment (see last paragraph in section 3.3). As we mentioned before, this ability is one of the signatures of explicit learning or explicit knowledge which may be a product of a qualitatively different system for learning and processing information compared to implicit learning that is usually thought to be involved in first language acquisition.

In the equal frequency condition, 11 out of 30 participants (~36%) were rule-staters, compared to 31 of 60 (~52%) in the minority default condition. The higher proportion of rule-staters in the minority default condition is likely due to the greater frequency of narrow rules, making them easier to learn. Figure 3 illustrates performance on novel stimuli in the equal frequency condition, comparing rule-staters and non-staters. Rule-staters performed near ceiling on the two narrow categories but significantly worse on the default category (Nasal vs. Elsewhere: odds-ratio=3.89, p < 0.001; -Ct vs. Elsewhere: odds-ratio=4.73, p < 0.001). This result aligns with our prediction that narrower categories would be learned more effectively than the broader default category. In contrast, non-staters exhibited a markedly different pattern, with performance hovering near chance, but still slightly above chance for the correct categories. This likely reflects the heterogeneity of the non-stater group, which includes both participants who failed to learn the pattern and those who implicitly detected some regularities.

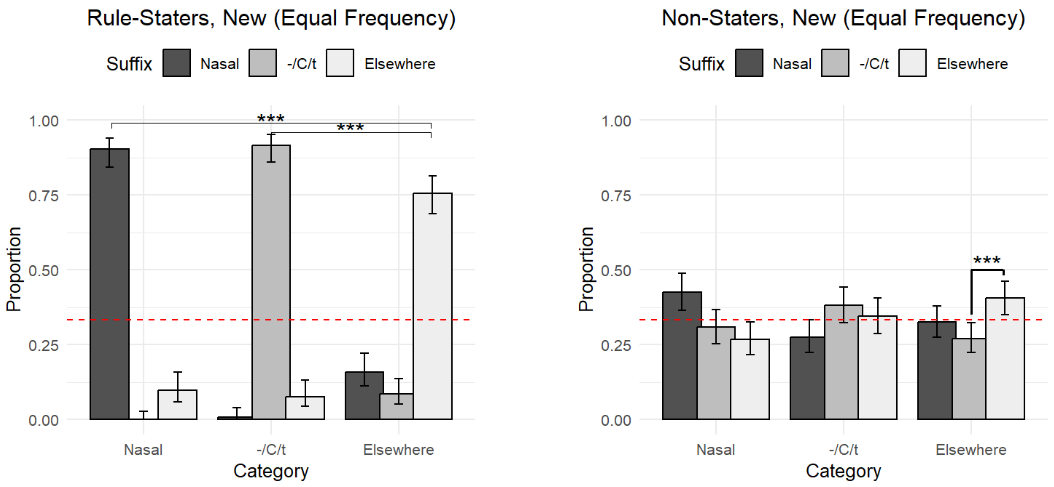


Figure 3. Performance in the Equal Frequency condition for rule-staters vs. non-staters. Proportion of correct responses (by category) for novel words.

Interestingly, in the minority default condition (see Figure 4), while the rule-staters look very similar to the rule-staters in the Equal Frequency condition, non-staters show a distinct pattern. In particular, non-staters perform significantly better on the most-frequent category compared to the mid-frequent category (odds ratio=1.54, p < 0.01) and the elsewhere category (odds-ratio=2.45, p < 0.001), confirming the effect of frequency for implicit learners. Additionally, they tend to generalize the most frequent suffix over the default suffix to novel words that follow the elsewhere distribution (odds-ratio=1.2, p=0.03). The second most frequent suffix was also used more frequently than the default, but not significantly so. This is consistent with our hypothesis 2(a) that high frequency would interfere with learning minority defaults. Thus, it appears that those participants who find and follow rules are better able to learn minority defaults, while participants who are not (yet) consciously aware of the pattern are influenced more by frequency than distribution when those two factors are in conflict.

Overall, we found that separating participants into rule-staters vs. non-staters led to a different picture of the results. While rule-staters appear to successfully generalize the default suffix to novel instances in both conditions, non-staters tend to overgeneralize the more frequent suffixes to novel words that have the elsewhere distribution. Although this preference was not very big, it suggests that there may be different learning biases for explicit vs. implicit learners or for learners at different stages in the learning process (assuming that explicit awareness at a later stage can develop from initial implicit learning, Sun et. al. 2005). Such an interpretation is consistent with Nevat et al. (2018) study, who found frequency to play a greater role early in the learning process and distribution to play a greater role at the end of learning.

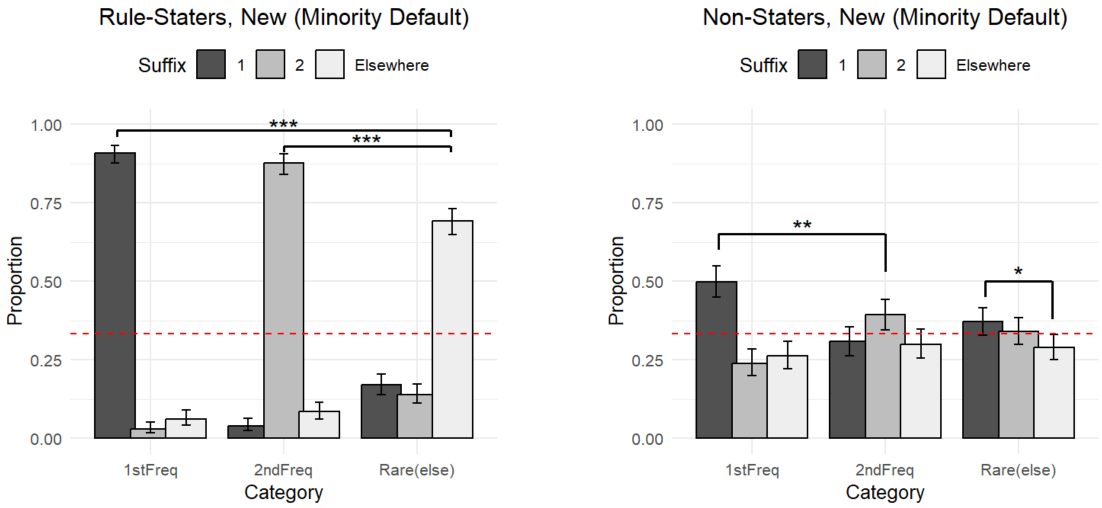


Figure 4. Performance in the Minority Default condition for rule-staters vs. non-staters. Proportion of correct responses (by category) for novel words.

4 Computational Modeling

Given the debate between rule-based vs. connectionist models of morpho-phonological learning about their ability to capture minority defaults, we test two computational models on our small artificial language dataset and compare the results to human learners. We use the Minimum Generalization Learner (MGL) as our rule-based model (see Section 4.1) and an encoder-decoder recurrent neural network (RNN) as our connectionist model (see Section 4.2).

4.1 *Minimum Generalization Learner* Minimum Generalization Learner (MGL) is a rule-based morpho-phonological learner created by Adam Albright and Bruce Hayes in the late 1990's and, among other things, used to model the English past tense, including modeling of human responses to a past-tense wug-test (Albright and Hayes, 2002). In more recent work Wilson and Li (2021) applied this algorithm to German, English and Dutch data from the Sigmorphon's UniMorph shared task and achieved competitive results.

The input to the algorithm are pairs of forms that stand in some morphological relation to each other (e.g., [mɪs]$_{\text{pres}}$, [mɪst]$_{\text{past}}$) and the output are phonological rules of the form: A → B / C __ D (A is realized as B in the context between B and D). A and B can be empty strings, and the environments C and D can be sequences of segments, feature bundles, and/or variables. The algorithm begins by creating word-specific rules first (e.g., ø → t/ mis __) and then works by recursively collapsing these rules to create more and more general rules. We refer the reader to Albright & Hayes (2002) for the specific description of the algorithm. Crucially, this learner generalizes by finding minimal differences between contexts in which the same morphological change occurs, and all intermediate rules are saved together with their confidence scores. A confidence score quantifies rule-accuracy adjusted for rule-scope. These scores are used to resolve competition among rules. That is, if multiple rules are applicable in the same context, the rule with the highest confidence will be chosen. Some of the rules produced by the model are redundant and can be pruned away. A rule R is redundant if it is a proper subset of another more general rule R' and R' has higher confidence score than R (guaranteeing that R will never apply).

Below we report the rules that the algorithm[2] found when trained on the same stimuli that the human subjects were trained on in our experiment. Note that the algorithm does not have a way to count syllables, and we did not include syllable feature in the training. The feature set we used in the training is included in the Appendix. First, consider the rules found at the end of learning in the Equal Frequency condition.

(1) Results of MGL in the equal-frequency condition. Non-redundant rules found by the learner with their confidence scores for each category.

1. -Ct category:
    1. ø → wa / X [-syllabic, -liquid, +ant] t __ # (0.93)
       "Insert suffix 'wa' after stems ending in an anterior consonant followed by t"
2. Nasal category:
    1. ø → jo / X [+syllabic, +voice, -diph] m __ (0.87)
       "Insert suffix 'jo' after stems ending in -Vm"
    2. ø → jo / X [+syllabic, +voice, -diph] n __ (0.85)
       "Insert suffix 'jo' after stems ending in -Vn"
3. Elsewhere category:
    1. ø → ler / X [-syll, -liquid, +cont, +ant]__ (0.85) (after [f, s, v, z, θ])
    2. ø → ler / X [-syl, -liquid, -voice, -COR]__ (0.78) (after [h,f,k,p]
    3. ø → ler / X [+syll, +voice, -diph]__ (0.78) (after vowels)
    4. ø → ler / X [-syll, -liquid, +COR, +ant]__ (0.72) (after [b, f, p, v])
    5. ø → ler / X [-syll, -liquid, +voice, +ant]__ (0.3) (after [b,d,v,n,z])
    6. ø → ler / X __ (0.29) (after any segment

You can see that for the narrow categories the learner finds one or two general rules that have high confidence. Unlike many of our participants, MGL learned a more specific rule for the -Ct category restricted to words ending in consonant clusters consisting of an anterior consonant followed by -t, not just words ending in -t. This is due to the "minimal" aspect of the MGL. That is, minimal generalization guarantees that more general rules are only created in the process of finding minimal differences between two more specific rules. Generalization to -t was not possible since there was only one specific -Ct rule. For the nasal category, the learner failed to find a single most general rule restricted to nasals. Instead, it found two separate rules for /m/ and /n/ when they are preceded by a vowel. It is not entirely clear why these two rules were not collapsed into a more general rule — perhaps more training data is required. Turning to the elsewhere category, we see that the learner abstracts away many specific rules whose disjunction "cobbles together" the heterogeneous default category. It does learn the most general rule that can be considered to be the default, "insert -ler after any segment," but this rule has a low confidence score and will only apply to those cases that are not already covered by the more specific rules for this suffix. Thus, this learner is able to predict that a default rule would apply across the board to any context that is not already covered by another rule. The results for the minority default condition are slightly different and are shown below.

(2) Results of MGL in the minority-default condition. Non-redundant rules found by the learner with their confidence scores for each category.

1. -Ct category:
    a. ø → wa / X [-syllabic, -liquid, +ant] t __ # (0.93)
       "Insert suffix 'wa' after stems ending in an anterior consonant followed by t"

[2] We used the implementation of the learner from this repository: https://github.com/colincwilson/MinimalGeneralizationLearner

2. Nasal category:
    a. ø → jo / X [-syllabic, -liquid, -cont, +voice] __ (0.95)
    "Insert suffix 'yo' after stems ending in voiced stops [b,d,g,n,m,ŋ]"
3. Elsewhere category:
    a. ø → ler / X [-syll, -liquid, -cont, -voic, -COR]__ (0.82) (after [k, p])
    b. ø → ler / X [[-syl, -liquid, +LAB, -COR, +ant]]__ (0.71) (after [b,f,p,v]
    c. ø → ler / X [+syll, +voice, -diph]__ (0.71) (after vowels)
    d. ø → ler / X [-syll, -liquid, -COR]__ (0.32) (after [b, f, g, h, k, m, p, v, ŋ])
    e. ø → ler / X [+voice]__ (0.169) (after voiced segments)
    f. ø → ler / X __ (0.163) (after any segment)

Note that in this condition the learner found one rule for the nasal category that is restricted to voiced stops, which is a more general category than nasals. However, this rule still has high confidence because no training data contradicted it: in the minority default condition, there were fewer words in the widely distributed "elsewhere" category and it so happened that none of them ended in voiced oral stops. Therefore in this condition, the learner will overgeneralize the nasal suffix to all voiced stops. The rules for the default suffix look similar in this condition to the rules found in the equal frequency condition, however the more general of these rules have lower confidence scores, reflecting the more narrow scope and greater number of exceptions characteristic of minority defaults.

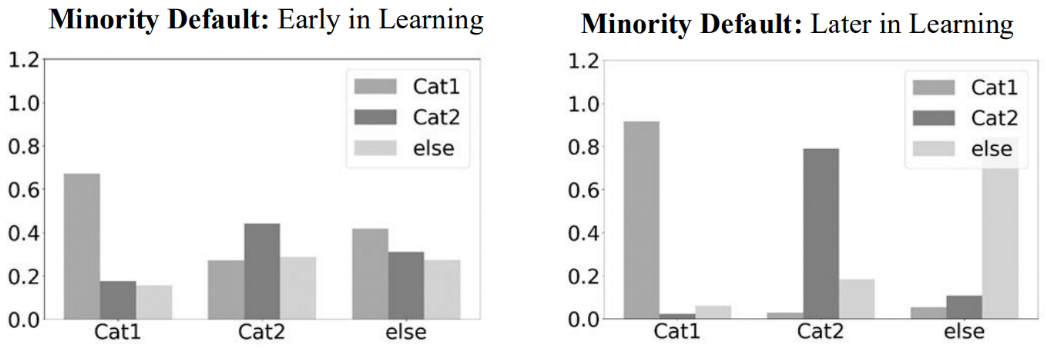
Overall, this learner will perform at ceiling on all stimuli, except minority default condition, in which it will overgeneralize the nasal suffix to all voiced stops, instead of just the nasal stops. One way in which this learner differs from experiment participants is that it generalizes based on the longest shared strings (starting right-to-left from the suffix), while the participants ultimately mainly paid attention to the last segment of the word (a simpler, less conservative hypothesis). This is evident from the fact that participants treated words ending in -Vt as belonging to the -Ct category, while MGL treats these words as belonging to the Elsewhere category.

Because MGL is transparent unlike the connectionist model we examine next, it allows us to see that defaults can be represented as disjunctions of multiple specific rules or as a general rule that has many exceptions (and hence low-confidence), but that can still apply when no other rule is applicable. The performance of rule-staters most closely resembles the predictions of this model.

4.2 *Encoder-Decoder* To test how an encoder-decoder neural network can capture the patterns used in our experiment, we implemented a model similar to ones used in past work on morphological learning and generalization (e.g., Kirov and Cotterell 2018, McCurdy et al. 2020). Our network was simpler than those that have been used on real-world languages, for two reasons: (1) it didn't need to be as complex to represent its inputs and outputs because our artificial language was considerably simpler and (2) a smaller model meant less computing power was necessary to perform our simulations of the experiment. In total, the model had 2 GRU layers each (Cho et al., 2014) in its encoder and decoder, and each of these layers had 10 nodes. All nodes used hyperbolic tangent activation functions.

The model was trained to map stems to one of the three suffixes (e.g., [bast] → [-wa]). The stems were represented as a sequence of numerical feature bundles (with standard phonological features and 1/0/-1 standing in for +/unmarked/-). Each output suffix was represented using a single timestep and a unique combination of two features. For example, [-wa] could be represented using the combination [1, 1] while [-ler] might be [1, 0]. The network was trained for 50 epochs, batch sizes of 1 (i.e., online learning), a learning rate of .0005, and 10 separate runs (with randomly sampled initial weights) in each condition.

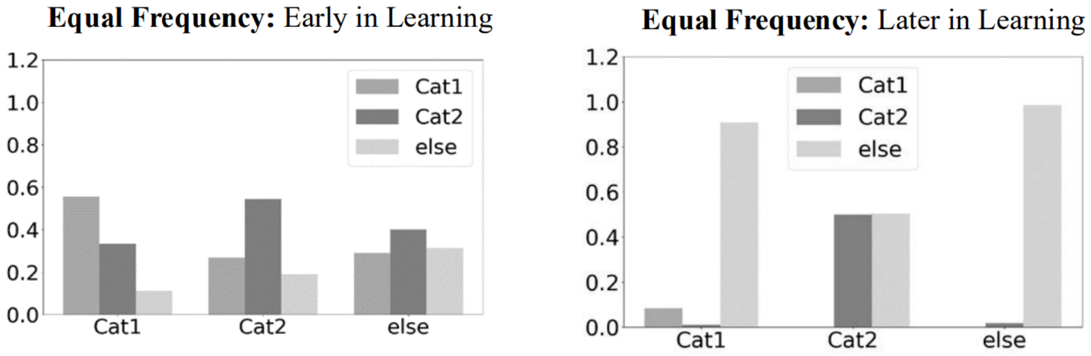
We trained and tested the model on the same kind of data as the human participants. Figure 5 shows performance of the ED model in the Minority default condition when tested on novel stems of one or two syllables long early in learning and at the end of learning. Early in learning, the models' performance looks similar to the performance of non-staters in Figure 4 (right panel). That is, there is a clear effect of frequency that interferes with learning the default. Later in learning (when the model had near-perfect accuracy), it successfully captured the patterns in the training data, with each kind of stem receiving the appropriate suffix majority of the time, similar to what we found with rule-staters.


Minority Default: Early in Learning

Minority Default: Later in Learning

Figure 5: The ED model's results in the minority default condition on novel data of the same type as training data. "Cat1" and "Cat2" correspond to Nasal and -Ct categories, respectively, with "else" corresponding to the elsewhere category.

However, the resemblance to human-performance breaks down in the equal frequency condition (Figure 6). In particular, while early in learning the model's performance is similar to the Minority Default condition, in the end the model fails to learn the pattern — it incorrectly overgeneralizes the default to the Nasal category and for the Ct category it is split between the -Ct-appropriate suffix and the default-suffix. Thus, in this condition, the model is biased towards the default in a way human participants are not.


Equal Frequency: Early in Learning

Equal Frequency: Later in Learning

Figure 6: The ED model's results in the equal frequency condition on novel data of the same type as training data. "Cat1" and "Cat2" correspond to Nasal and -Ct categories, respectively, with "else" corresponding to the elsewhere category.

We tested the ED model's performance on three-syllable words separately to see if the model ignored syllable number in the same way that our participants did (see Figure 7). What we found is that in both conditions the model performed similarly to humans (and differently from the way it performed on 1 and 2 syllable words). Namely, it treated three-syllable words that ended in nasals or -Ct as belonging to the narrow categories (even though in the training those words were never 3 syllables long). However, the number of three-syllable words ending in -Ct or nasal consonants was small, which may limit the strength of this observation. Interestingly, the model did not have the same problems in the equal frequency condition as it did when tested on 1 and 2 syllable words. Like human participants, the ED model appeared to generalize suffix selection based primarily on the final segment of the word. However, because in the Equal Frequency condition it performed differently on three-syllable words compared to the shorter words, we cannot say that it completely ignored syllable count.

Overall, the modeling results suggest that this implementation of the encoder-decoder model did not fully match human performance. Surprisingly, the model did better in the minority-default condition compared to the equal frequency condition, where it failed to learn the elsewhere distribution. One reason could be that when the suffixes had equal frequency, the default was actually learned faster and assumed to apply to a wider set of contexts.

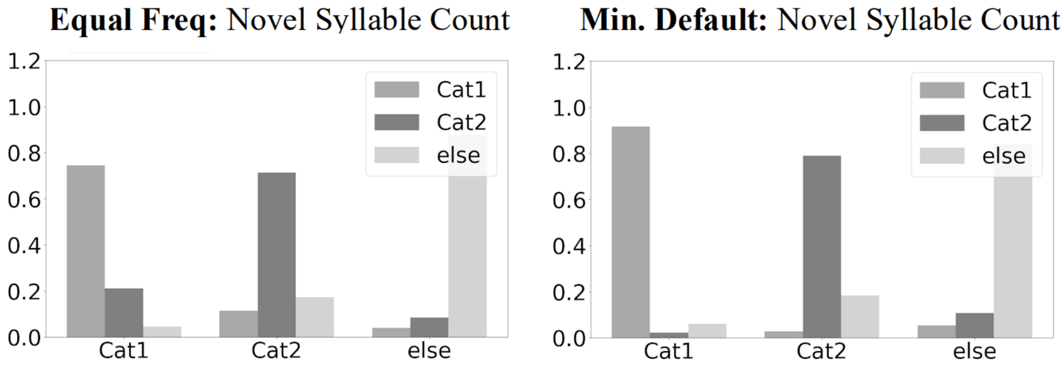


Figure 7: The ED model's results in both conditions on novel 3-syllable words

5 Conclusions

In this study, we examined how frequency and distribution shape the learning of default patterns in an artificial language, contributing to the broader debate on rule-based vs. connectionist models of morpho-phonological learning. Our results show that human participants successfully learned the default pattern, even when it was a minority, suggesting that distributional cues can override frequency under certain conditions. Similarly, both a rule-based and a connectionist encoder-decoder (ED) model were able to learn the minority default pattern, challenging the claim that connectionist models necessarily fail when generalization is not frequency-driven (Pinker & Ullman, 2002).

However, we found that frequency did interfere with learning minority defaults, particularly in the early stages of learning. The ED model exhibited this effect most clearly, and our results suggest that human learners may also rely on frequency early on, as those who did not explicitly verbalize a rule tended to overgeneralize the most frequent suffix. On other hand, rule-staters were more successful at learning the default distribution. This supports the idea that rule discovery plays a role in overriding frequency biases.

The poor performance of the ED model in the equal-frequency condition was unexpected, raising questions about how such models represent defaults. Further investigation is needed to determine whether architectural modifications—such as attention mechanisms or explicit inductive biases—would allow the model to better approximate human performance.

As for the rule-based Minimum Generalization Learner (MGL), our results suggest that defaults may be internally represented as a disjunction of multiple specific rules. This would explain why defaults, particularly minority defaults, are harder to learn. At the same time, MGL was able to derive a fully general default rule that applies in any context where no more specific rule is available. Unfortunately, we were unable to directly test this property in human learners, as our experimental stimuli were not sufficiently distinct from training examples. To address this limitation, we are currently conducting a follow-up study designed to examine whether human learners apply the default suffix in truly novel contexts.

6 References

Albright, Adam and Bruce Hayes. (2002). Modeling English past tense intuitions with minimal generalization. In

Proceedings of the ACL-02 workshop on Morphological and phonological learning (pp. 58-69).

Alshboul, Sabri S., Yousef M. Al-Shaboul and Sahail M. Asassfeh. (2012). The Elsewhere Inflection: Evidence from Nominal Patterns in Modern Standard Arabic. *SKASE Journal of Theoretical Linguistics*, *9*(1).

Boudelaa, Sami and M. Gareth Gaskell. (2002). A re-examination of the default system for Arabic plurals. *Language and cognitive processes*, *17*(3), 321-343.

Corkery, Maria, Yevgen Matusevych and Sharon Goldwater. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. *ArXiv Preprint ArXiv:1906.01280*.

Flannagan, Michael. J., Lisbeth S. Fried and Keith J. Holyoak. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(2), 241.

Gawlitzek-Maiwald, Ira. (1994). How do children cope with variation in the input? The case of German plurals and compounding. In R. Tracey (ed.), How Tolerant is Universal Grammar? Essays on Language Learnability and Language Variation (Tuebingen: Niemeyer), 225-66.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, *41*(4), 313-360.

Hare, Mary, Jeffrey Elman and Kim Daugherty. (1995). Default generalisation in connectionist networks. *Language and cognitive processes*, *10*(6), 601-630.

Kirov, Christo and Ryan Cotterell. (2018). Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker & Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Marcus, G. F., U. Brinkmann, H. Clahsen, R. Wiese, & S. Pinker. (1995). German Inflection: The Exception That Proves the Rule. *Cognitive Psychology, 29*(3), 189-256.

MacWhinney, Brian. (1978). The acquisition of morphophonology. Monographs of the Society for Research in Child Development, 43.

McCurdy, Kate, Sharon Goldwater and Adam Lopez. (2020). Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1745-1756).

Nevat, Michael, Michael T. Ullman, Zohar Eviatar and Tali Bitan. (2018). The role of distributional factors in learning and generalising affixal plural inflection: An artificial language study. *Language, Cognition and Neuroscience*, *33*(9), 1184-1204.

Park, Tschang-Zin. (1978). Plurals in child speech. Journal of Child Language, 5, 237-50.

Pinker, Steven and Alan Prince. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

Pinker, Steven and Michael T. Ullman. (2002). The past and future of the past tense. Trends in cognitive *sciences*, *6*(11), 456-463.

Prickett, Brandon, Traylor, Aaron, & Pater, Joe. (2022). Learning reduplication with a neural network that lacks explicit variables. Journal of Language Modelling, 10(1), 1-38.

Sun, Ron, Slusarz, Paul, and Chris Terry. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological review*, *112*(1), 159.

Sutskever, Ilya, Vinyals, Oral, and Quoc V Le. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104–3112.

Wang, Yuxuan, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135.

Wilson, Colin, and Yane S.Y. Li. (2021). Were we there already? Applying minimal generalization to the SIGMORPHON-UniMorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 283-291).

Zaretsky, E., Müller, H. H., & Lange, B. P. (2016). No default plural marker in Modern High German. *Italian Journal of Linguistics*, *28*(1), 1-28.