Interactions Between Linear Order and Lexical Distributions in Artificial Language Learning

Holly Jenkins (holly.jenkins@education.ox.ac.uk)

Department of Education, University of Oxford

Michael Ramscar (michael.ramscar@uni-tuebingen.de)

Department of Psychology, University of Tübingen

Elizabeth Wonnacott (elizabeth.wonnacott@education.ox.ac.uk)

Department of Education, University of Oxford

Abstract

How do children learn the appropriate scope of linguistic generalizations? One proposal is that prediction error and cue competition enable them to implicitly reduce their uncertainty about the various cues to linguistic patterns. Previous work has employed artificial language studies to test the predictions of error-driven models against the performance of (adult) human participants (Ramscar et al., 2010). A critical prediction of these models - that linear relations between linguistic and environmental cues can critically affect generalization - has received much empirical support. For example, Vujovic et al. (2021) found that suffixing languages supported the learning of discriminating cues, and overgeneralization avoidance, better than equivalent prefixing languages. The current study addresses a limitation of previous studies: the use of unnatural flat distributions, which contrast to the skewed distributions ubiquitous in natural language. Although some of our results are consistent with model predictions, there were divergences. Possible reasons for these are discussed.

Keywords: Language learning; Lexical distributions; Order effects; Morphology; Affixation

Introduction

How do children learn to recognize consistent linguistic patterns and then generalise them to new utterances? Traditional theories of language acquisition often assume that children's mastery of these patterns relies on abstract symbolic rules, and that these rules are innate components of the language processing system. By contrast, an alternative explanation proposes that children rely on domain-general learning mechanisms to acquire these patterns, and that they use these mechanisms to detect and extract probabilistic structures that are implicitly present in the world and the linguistic environment.

For example, it has been proposed that children make use of prediction error and cue competition to reduce uncertainty about the systematic cues to linguistic patterns, and rely on these cues to generalize their use to new contexts. This leads to the prediction that linear order matters in learning: appropriate generalization over cues requires those cues to proceed the relevant linguistic patterns as outcomes. Ramscar et al. (2010) tested this prediction in an artificial language study where participants learned labels for novel objects either in a word-picture order or a picture-word order. Participants showed more appropriate generalization - i.e. learning the correct scope of a particular set of visual features as cues to a label - in the picture-first condition, i.e. the condition in which cue competition over the features had been possible.

The core finding from Ramscar et al. (2010) has since been replicated in a number of different domains (see Viviani et al, 2024 for a review). In particular, the discriminative nature of error-driven learning suggests that learning differs based on whether an affix is presented before (prefix) or after (suffix) a complex stem. Critically, suffixes are thought to promote greater cue competition, which constrains generalisation through discriminative learning.

To test this empirically, Vujović et al. (2021) conducted a series of experiments measuring the learning of inflectional morphology with prefixes and suffixes. In these experiments, participants were aurally exposed to examples of referents from an artificial language consisting of four classes of nouns, but only two affixes which were associated with varying semantic and phonological cues (see Fig. 3). Therefore, learning the appropriate cues to an affix relies on identifying a key set of discriminating features (circled in Fig. 3) and, critically, dissociating a frequent and salient, but uninformative cue (the body colour & shape) from less frequent but informative cues to category membership. In the training phase, the audio labels were presented in such a way that in both the prefix and suffix condition, the image was only displayed on screen for the duration of the noun. Therefore, in the prefix condition, participants heard the prefix before seeing an image of the associated referent (mimicking the fact that in natural language processing, the identity of the noun is not known until it is heard). In the

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

suffix condition, participants had already seen the referent image associated with the suffix. The conditions were designed so that the images are displayed on the screen for the same duration (see Fig. 3); however, cue competition was only possible in the suffix condition. Following training, participants were tested on a number of post-exposure learning tasks. Included in these tasks was a generalisation task, where participants were presented with novel visual stimuli and instructed to choose between two possible audio labels based on the noun's semantic and phonological features.

Computational models designed to simulate the training phase indicated that in the prefix condition, the high-frequency but non-discriminating features were not distinguishable from the low-frequency but discriminating features. However, the suffix condition was not as affected by frequency, as cue competition allows the model to learn that the high frequency feature has a lower predictive value than the lower frequency, but informative feature. Vujović et al. (2021) confirmed the key predictions from these simulations in a large-scale web-based artificial language learning experiment, demonstrating that participants learning prefixes were more likely to show incorrect reliance on frequent, but uninformative cues, compared to the suffix condition. This was clearly seen as greater over-generalization with "minority" noun classes in the prefix condition indicating that participants were unable to discriminate between the various cues.

A limitation of Vujović et al. (2021) (and the original study by Ramscar and subsequent replications) is that the artificial languages were based on a flat input distribution where all of the nouns were presented equally frequently in training (although the subclasses of fribbles differed in frequency; see Fig. 1). This is vastly different to natural language, where there is extensive evidence that exemplars of categories consistently follow highly skewed distributions: some individual items are exponentially more likely to occur than others (Estoup, 1916; Zipf, 1946; Linke & Ramscar, 2020). Evidence from previous artificial language learning experiments suggests that these skewed distributions lead to different patterns of learning that flat distributions (Casenhiser & Goldberg, 2005; Wonnacott et al., 2017; Wolters et al, 2024). It is currently unclear how implementing more realistic exponential distributions will interact with previous manipulations of temporal order in prefix and suffix learning.

Therefore, in the current work we ask: What are the consequences of employing more realistic exponential distributions, and how will this interact with our manipulations of temporal order in suffix and prefix learning? We first present computer simulations to determine whether different patterns of learning with two types of skewed input distributions. Both used exponential distributions, but differed in terms of which items were selected to be the highest frequency subset. We then aim to first replicate the findings of Vujović et al. (2021) with a flat input distribution, before testing the predictions of the current simulations in several artificial language learning experiments.

Computational Simulations

We created three input sets in which - as in Vujovic et al. (2021) - each of the two affixes is represented as a single discrete feature while nouns and their referents (the fribbles) are represented as a set of features. The vowel is the discriminating feature of the stem, and this was coded as a discrete phoneme (Ramscar et al., 2010). The semantic features that are relevant for generalisation were also coded as discrete features. Two additional features were also coded to represent features that were unique to an individual item. A context cue (the same across all trials) was also included. The models were trained to asymptote (8000 trials).

In each of the three input sets the same nouns are associated with the same affixes- and in each case there is a minority subclass of fribbles associated with each affix. What differs is the frequency of each noun (Fig. 1). In the flat input (identical to the original experiment), frequencies are equal. By contrast, in the exponential distribution, type and token frequency come in conflict, since the minority subclass has the highest token frequency. In the balanced exponential distribution, the minority items have the same token frequency (same total number of presentations) even though there are fewer exemplars.

The design of the simulations was identical to Vujović et al. (2021). For each type of input, a prefix and suffix version of the model was run: In the prefix model, a single cue (the affix) was used to predict a set of outcomes (the phonological and semantic features of the referent). This was the opposite in the suffix model: the features were the cues and the affix was the outcome.

Simulation Results

Generalisation was determined by training the models to asymptote and looking at the weights between features relevant for generalisation (in this case, body shape, the discriminating feature, and the vowel) and each of the affixes, and finding the sum of the associative strengths. A model that can successfully discriminate between the informative and uninformative cues will have the largest weights for the informative cues and the lowest weights for the uninformative cues.

The results are shown in Figure 1, which shows the weights for the affix "*ma*" (although these weights would mirror those for the affix "*ge*"). The red and blue bars show the weights for body-shape/colour (these are uninformative, and not consistently associated with the affix) and the grey bars show the discriminating semantic cues (these are the cues consistently associated with one of the four subclasses of fribbles (Fig. 2), and thus consistently associated with one of the affixes. For the flat distribution, we see in the prefix condition, while the red body is highly associated with the affix (due to its frequency of occurrence with that affix), this

difference is present but much less marked in the suffix condition. In contrast, only in the suffix condition has the model developed negative weights for cues d1 and d2 and for the contrasting vowel sound, which are cues that consistently occur with the other affix ("ge"). This leads to the prediction of an interaction between type-frequency (minority-class (LTF) / majority-class (HTF)) and affix condition (prefix/suffix) where the prefix condition are expected to show a stronger effect of frequency than the suffix condition, and to do badly in categorizing minority category items (since they tend to associate them with the other affix based on body shape and to pay less attention to the presence of the wrong discriminating features).

For the exponential distribution, when the minority-class items have the highest token frequency, the model still predicts that for the prefix condition, there will be an interaction between affix and type frequency, but in the reverse direction since it is driven by the strong association of the blue body with the affix (i.e., a token frequency effect). Interestingly, the suffix model is nearly identical to when the distribution is flat, i.e. predicting a small type frequency effect and strong learning of the discriminating features. Finally, for the balanced exponential distribution, the model predicts that for the prefix condition the effect of type frequency will be made redundant when the token frequency is balanced across the minority and majority classes (no type or token frequency effect). Once again, the model predicts no impact of input distribution on the suffix condition. Generally speaking, the models suggest that in the prefix condition, performance appears to reflect the token frequency of the items (i.e., the input distribution), but the suffix condition appears to show a small effect of type frequency that is immune to changes in token frequency. Critically, this involves the development of negative weights for certain discriminative features, which allows for improved performance in classifying minority-class items.

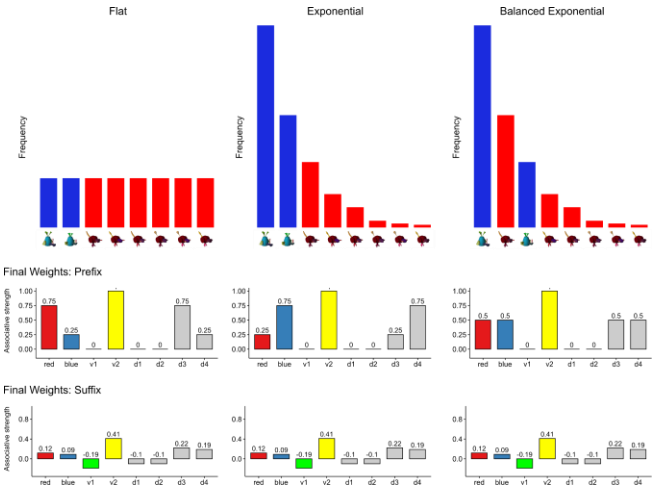

Figure 1. Distribution of fribbles which occur with affix "ma" in each of the three input sets: flat (as in Vujovic et al., 2021) exponential and balanced exponential. The distribution will be mirrored for the alternative affix ("ge"), with output weights from the computational simulations for each distribution

Method

Participants

All participants were recruited via Prolific and completed the experiment on Gorilla Experiment Builder. For the flat input distribution (replication of Vujović et al., 2021), 100 participants (42 female, 57 male, 1 who did not disclose, mean age = 31.04, SE = 0.66) were recruited. For the exponential distribution, 80 participants (38 female, 42 male, mean age = 32.73, SE = 0.83) were recruited, while 80 participants (35 female, 45 male, mean age = 32.09, SE = 0.69) completed the experiment with the balanced exponential distribution.

Stimuli

The same stimuli were used as in Vujović et al. (2021). The stimuli were split into two categories. For the visual stimuli, each category contained six HTF and two LTF items. The audio stimuli consisted of 16 training and 16 test nouns, each comprising of a CVC syllable. Half the training and test items were allocated to Category 1, and contained the vowel sound /u:/, and the Category 2 stimuli contained the vowel sound was /i:/. To improve on the audio quality from those previously recorded, the 32 audio stimuli were reproduced using Microsoft Azure AI Speech Studio (United Kingdom English, "Sonia").

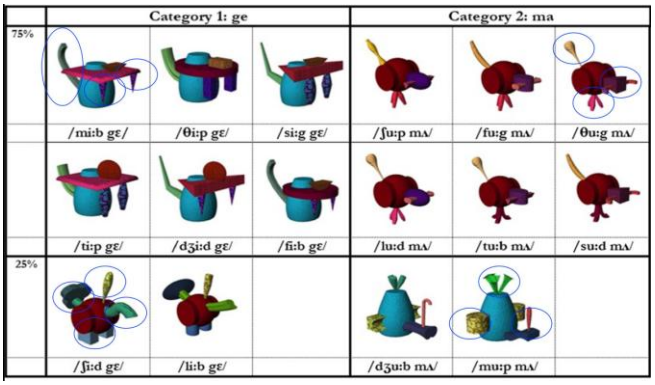

Figure 2. Sample training set for the suffix condition. Discriminating features are circled.

Procedure

The procedure was identical to Vujović et al. (2021). Participants were randomly assigned to complete either the "suffix" or "prefix" condition, and completed a training phase, followed by a number of post-exposure tasks aimed to determine whether learning had occurred. For brevity, this paper will only discuss the findings from the generalisation task.

Training Phase

The training phase lasted approximately 15 minutes, consisting of a total of 256 trials split into 4 blocks of 64 trials, each separated by a break. In the prefix condition, in each trial, the affix was played first (225ms), immediately followed by the noun (350ms). The corresponding image was displayed while the noun was presented and remained on screen for an additional 450ms after. In the suffix condition, the image was displayed for 450ms, after which the noun was played, followed immediately by the affix. The image was not displayed on screen for the duration of the affix in either condition.

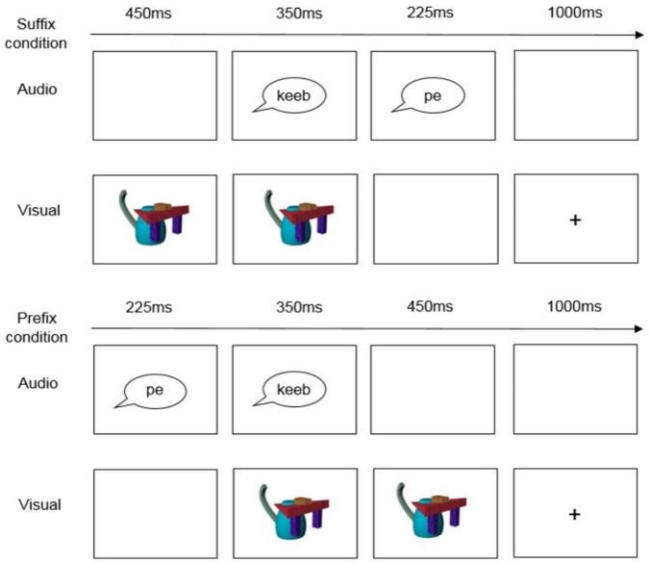

Figure 3. Schematic representation of a single training trial in each condition. Figure taken from Vujović et al. (2021).

Generalisation Task

This task tested whether participants were able to generalise the learning that occurred in the training phase to novel stimuli. Participants were presented with an image of a novel fribble belonging to one of the categories. After 500ms, a blank speech bubble was displayed in the lower left corner of the screen while the first novel audio was played. After 500ms, the second novel audio was played while an empty speech bubble was displayed in the lower right corner of the screen. After 500ms, both speech bubbles were redisplayed on the screen, and the participants were tasked with clicking the speech bubble that matched the audio corresponding to the image displayed on screen. The order of presentation of the target and foil audios was randomised. The foil audio was from the same category as the target (containing the same vowel sound) but was paired with the affix from the opposite category. For example, if the target image was from Category 1, the target audio was *foop ge* (vowel A + affix A) and the foil audio was *kood ma* (vowel A + affix B). Each novel item appeared as the target once, totalling 16 trials.

Data Analysis

The data from the generalisation task will be analysed using GLMs (logistic regression with a binary distribution), as the accuracy data are binary. Affix (prefix versus suffix) and type frequency (high versus low) were included as fixed effects. (Note that type frequency is matched across the input sets; if there are effects of token frequency this will be seen in a reversal (exponential) / removal (balanced exponential) of the type frequency effect. Participant was included as a random effect with a random intercept by-participants slope for type frequency. All fixed factors were centred so that the intercept reflected the grand mean (and coefficients are main rather than simple effects). Distribution was included as a fixed factor with 3 levels (*flat*, *exponential* and *balanced exponential*), and coded using a simple coding with *flat* as the reference level. This allows us to inspect the contrast between this condition and the other conditions and ensures that the intercept continues to reflect the grandmean.

Results

We included data from all three distributions in the analysis (output shown in Table 1). Overall, there was a main effect of type frequency (β = 0.432, SE = 0.081, z = 5.313, p < .001), which was not predicted by the simulations. Across the experiments, humans seem to show improved performance for majority-class items over minority class items even with varying token frequency, despite the models suggesting that performance would be affected by token frequency in the prefix condition.

Table 1. Model output. Note that simple effect coding means that the intercept reflects the grandmean.

	Estimate	Std. Error	z value	Pr(>|z|)
intercept	0.416	0.053	7.852	< 0.001
affix	0.005	0.105	0.043	0.966
type frequency	0.432	0.081	5.313	< 0.001
flat vs exponential	-0.126	0.126	-0.996	0.319
flat vs balanced exponential	0.103	0.126	0.815	0.415
affix * type frequency	-0.395	0.160	-2.464	0.014
affix * flat vs exponential	-0.149	0.253	-0.589	0.556
affix * flat vs balanced exponential	0.055	0.253	0.217	0.828
type frequency * flat vs exponential	-0.596	0.193	-3.090	0.002
type frequency * flat vs balanced exponential	-0.293	0.193	-1.516	0.129
affix * type frequency * flat vs exponential	0.231	0.385	0.601	0.548
affix * type frequency * flat vs balanced exponential	0.277	0.387	0.716	0.474

Across the three types of distribution, there was also an interaction between affix and type frequency (β = -0.395, SE = 0.160, z = -2.464, p = .014). When examining this interaction in more detail, analyses indicated that this was driven by overall above chance performance in the minority-class items for the suffix condition only (β = 0.300, SE = 0.094, z = 3.209, p = 0.001), whereas there was no evidence that the prefix condition performed above chance in these items (β = 0.098, SE = 0.093, z = 1.053, p = 0.292). This is consistent with the suffix condition being less affected by token frequency, thus enabling participants to be better at

rejecting the incorrect affix despite the association with the body shape/colour.

There was also an interaction between type frequency and the flat and exponential distributions (β = -0.596, SE = 0.193, z = -3.090, p = 0.002), reflecting that the improved performance on majority-class items that is found with the flat distribution is no longer present with the exponential distribution. Recall that the output of the simulations suggested that in the prefix condition, the type frequency effect would be reversed with the exponential distribution, and while this was not the case, the effect of type frequency is weakened (though not specifically in the prefix condition)

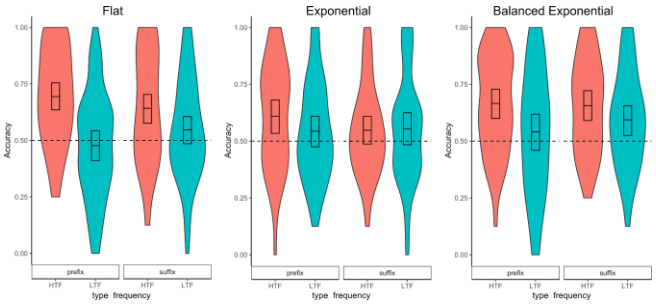

Figure 4. Performance in the Generalisation tasks for flat, exponential and balanced exponential distributions. The boxes represent CI around the mean. Chance performance (50%) is indicated by the dashed lines.

Discussion

Our simulations suggested that the learning of type frequency by participants in the prefix condition would be affected by the token frequency of the items, but there would be a small effect of type frequency in the suffix condition that would not be affected by changing the token frequency. When testing the predictions of the models in adult humans in an artificial language learning experiment, we found an overall interaction between type frequency and affix. This pattern of performance suggests that in the prefix conditions, there is a general tendency to incorrectly over-associate the minority-class items with the wrong affix, leading to poorer performance in these items relative to the suffix condition. In the suffix condition, our results further suggest that participants appear to have learned to negatively weigh certain incorrect discriminating features, indicating that they had better identified the features that enabled them to correctly classify the minority-class items.

However, the model suggested that an exponential input distribution, with minority-class items having the highest token frequency, would indicate that token frequency was driving learning in the prefix condition. The model suggested a reversal in the direction of the interaction for the prefix condition, in that participants would now show improved performance on the more frequent minority-class items. However, the data from our participants instead indicates better performance for the majority-class items across both prefix and suffix conditions, albeit with a weaker effect. Finally, while the models predicted that a balanced exponential distribution (where type and token frequency were balanced) would negate this interaction, human data still shows an effect of type frequency, indicated by better performance for majority-class items.

Overall, the findings from our experiments with humans indicate that humans seem to be more sensitive to type frequency than token frequency, as we consistently found evidence of better performance in majority-class items over minority-class items, even when the model predicted poorer performance based on changing token frequency. In this respect, our data suggests that human behaviour in both conditions matches most closely with the models learning in the suffix condition.

Given the previous findings of Ramscar et al. (2010), and subsequent replications, why does the learning of our participants not fully reflect the impact of linear order seen in the model? In particular, why are participants in the prefix condition not strongly impacted by token frequency as the model predicts? One, albeit tentative, explanation concerns the way in which our participants approach our learning task: While the models reflect purely implicit learning processes thought to be relevant in language acquisition, at least some of our adult participants are likely strategizing and actively trying to work out the system relating noun features to affixes. For example, although the paradigm was designed to tightly control the presentation of the stimuli, so that the information is ordered in time (in line with the presentation to the model) our adult learners were clearly capable of using memory processes to reorganise the information as they encountered it (e.g., having seen a noun, actively map it to the affix, regardless of when that affix occurred). This would serve to weaken the ordering manipulation, and may encourage more suffix-like learning, leading to benefits of variability (type frequency) over token frequency. While we acknowledge that this suggestion is speculative, it is consistent with a large body of work suggesting that adult language learning can be impacted by explicit hypothesis formation (e.g. DeKeyser, 2000), which may underpin weaker ultimate learning outcomes (Johnson & Newport, 1989; Hartshorn, 2018).

References

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. Developmental science, 8(6), 500-508

Estoup, J. B. (1916). Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse. Institut sténographique.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. Studies in second language acquisition, 22(4), 499-533

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. Cognition, 177, 263-277.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. Cognitive psychology, 21(1), 60-99.

Linke, M., & Ramscar, M. (2020). How the probabilistic structure of grammatical context shapes speech. Entropy, 22(1), 90

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. Cognitive science, 34(6), 909-957.

Viviani, E., Ramscar, M., & Wonnacott, E. (2024). The Effects of Linear Order in Category Learning: Some Replications of Ramscar et al.(2010) and Their Implications for Replicating Training Studies. Cognitive Science, 48(5), e13445.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. Journal of Memory and Language, 119, 104231.

Wolters, L., Lavi-Rotbain, O., & Arnon, I. (2024). Zipfian distributions facilitate children's learning of novel word-referent mappings. Cognition, 253, 105932.

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. Journal of Memory and Language, 95, 36-48

Zipf, G. K. (1946). The psychology of language. In Encyclopedia of psychology (pp. 332-341). Philosophical Library.