

The effects of type and token frequency on generalization

Zachary Nicholas Houghton^{1,2}, Zara Harmon³, Vsevolod Kapatsinski¹

¹University of Oregon, ²Vail Systems, ³Max Plank Institute for Psycholinguistics

Language is productive, and as a consequence the crucial task of a language learner is to generalize the use of a form beyond the contexts in which it has been encountered, to novel contexts [1,2,4]. Linguistic productivity has been demonstrated to be driven by type and token frequency of the form [3]. Specifically, recent work suggests that both high type frequency and high token frequency lead to increased probability of a form being used in a novel context [3].

Whereas linguistic productivity generally refers to extending a form to new formal contexts, e.g., extending an affix to new stems, it is also often important to extend a form to new semantic contexts. That is, it is important to be able to generalize a word's *meaning* beyond the context it is learned in. For example, one has to learn that the one-eyed, pirate cat that wanders the neighborhood can still be referred to by the form *cat*, despite the fact that a learner may never have encountered a one-eyed cat before.

It is unclear, however, whether generalization across formal and semantic contexts is driven by the same factors. The present study investigates this by deploying an artificial language learning experiment in which type and token frequency of different suffixes are manipulated (Figure 1). Specifically, two suffixes are taught to correspond to the “big plural” (several big creatures) meaning, while two alternative suffixes are taught to correspond to the “diminutive singular” (a single small creature) meaning. In training, each of these suffixes is paired with various creatures. We manipulate the number of times each suffix occurs with each stem (a token-frequency manipulation) as well as the number of different *types* of creatures each suffix occurs with (a type-frequency manipulation). In testing, we examine how learners extend these suffixes to a novel meaning (diminutive plural) which is partially compatible with each suffix. Additionally, some creatures are seen in training while others are not, thus contexts can be novel in stem or novel in meaning. We tested participants in production and forced form-choice.

We analyzed the results using Bayesian mixed-effects regression models with weak, uninformative priors. For the production and forced-choice tasks, the dependent variable was whether participants produced/selected the frequent suffix for each stem and meaning condition. The statistical analyses revealed meaningful effects for novel stem and meaningful interaction effects between novel stem and type frequency as well as token frequency and original meaning in the production task. In the 2AFC task, the only meaningful effect was a small interaction effect between type frequency and novel meaning. In other words, when there was a high type frequency participants produced the novel meaning more for novel stems, while for high token frequency, they produced the original meaning more (regardless of stem type). Further, when both forms were made accessible, this effect disappeared.

Our results suggest that type and token frequency influence both semantic extension and productivity / extension to new stems, suggesting that both semantic extension and linguistic productivity are potentially driven by the same underlying mechanism (Figures 2 and 3). This conclusion is reinforced by the finding that the data is well accounted for by a simple logistic perceptron model [4, 5] (Figure 4). Specifically, in production type frequency leads to semantic extension for novel stems, while token frequency increases probability of using that suffix to describe the original meaning (regardless of stem). Increased type and token frequency together lead to an increased probability of using the suffix regardless of stem or meaning type. The perceptron model also sheds light on the cues that participants use to perform the task: the model performs best when semantic cues are more salient than phonological cues. However, both phonological and semantic cues are necessary, suggesting that the choice of a form in a novel context is driven jointly by semantic and phonological features of that context, and the extent to which the discriminative features of the known contexts in which the form has appeared also appear in this novel context.

References

[1] Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, 114(2), 245.
[2] Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2-3), 150-177.
[3] Jarosz, G., Hughes, C., Lamont, A., Prickett, B., Baird, M., Kim, S., & Nelson, M. (2025). Type and token frequency jointly drive learning of morphology. *Journal of Memory and Language*, 104666.
[4] Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.
[5] 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*, 38(4), 575-596.

Figures and Tables

