Understanding Morphological Acquisition via Distributional Learning Models

Introduction. The question of how a child learns her native language remains highly debated in linguistic research. Cross-linguistically, children learn much of the morphology (word structure) of their native language by age three, when their vocabulary is approximately 1000 words.[1] Yet the frequencies of words in child-directed speech follow a skewed, roughly Zipfian distribution, with the frequency of a word being proportional to the inverse of its rank.[2] This means that a few words may occur hundreds of times in the child's linguistic input, but most occur only a few times. Similarly, most words only appear in a few of their possible inflected forms (e.g. the child may hear *fall* and *fell*, but not *falls* or *fallen*).[3] Such input contrasts with the larger, more saturated data used by most machine learning systems today. How, then, does a child learn the morphology of her native language from such a skewed, sparse input? **During my PhD, I seek to answer this question via computational modeling of morphological acquisition.**

A plausible model of morphological acquisition should follow the developmental and behavioral patterns of children, which may be studied experimentally and through analysis of errors in children's speech. *Productivity* of a linguistic process is marked by its ability to generalize to novel contexts, and is a foundational component of language. In the famous *wug* study, for example, most English-learning children generalized *-ed*, *-ing* and *-s* to novel verbs (e.g. *gling*) by age three, demonstrating that they had learned the productivity of these suffixes.[4] Further, most errors in children's speech are caused by over-application of productive processes: English verb acquisition is known to follow a "U-shaped" curve, where a sudden dip in performance is caused by the overapplication of *-ed* to irregular verbs (e.g. *goed*, *feeled*) when the child discovers its productivity.[5] Several promising computational models that account for these facts utilize the Tolerance Principle (TP), a threshold of productivity which posits that a child generalizes a process when it is more computationally efficient to do so under a Zipfian distribution.[2] Such distributional learning models have been the focus of my undergraduate research: working with Dr. Charles Yang, I developed a model that acquires meaning-form mappings (e.g. PAST = *-ed*) between suffixes and their corresponding semantic features, such as person (first, second or third), number (e.g. singular, plural), and tense (e.g. past, present, future).[6] This model follows developmental patterns and correctly acquires morphological rules on small vocabularies of Spanish and English verbs. I also created a model that acquires such mappings for German plural nouns and English verbs, even displaying U-shaped regression in English, and contributed to a third model with comparable results.[7] **It is my goal to build on these models to create an integrated, incremental, and cognitively-plausible model of morphological acquisition that succeeds on a wide array of languages.**

Aim 1: To create a model of incremental morphological acquisition that succeeds on a typologically diverse set of languages. While the models outlined above provide promising results, no single model is able to account for all languages: the latter two succeed on concatenative, non-agglutinative languages (e.g. English, German), but fail to model non-concatenative (e.g. Hebrew, Arabic) and agglutinative (e.g. Spanish, Swahili, Japanese) languages. Segmenting a word into morphemes is more challenging in such languages: in Spanish, for example, a verb may take multiple suffixes (e.g. *ama-ba-s* = love-past-2nd+singular = "you loved"). These models are also not incremental learners, but extract morphological rules from fixed-size vocabularies; this contrasts with the incremental nature of language acquisition. In my first aim, I thus plan to build on the models above to design a novel algorithm that incrementally acquires morphological rules across agglutinative, non-agglutinative, and non-concatenative languages. While current models take in each item as a lemma, inflected form, and semantic feature set (e.g. *walk*, *walked*, {PAST}), the child may be able to group each of the inflected forms in which she has seen a word (e.g. *get*, *gets*, *gotten*). I hypothesize that doing so will allow for cleaner segmentation and identification of morphemes, helping the learner to succeed on a wider array of languages. To test this, I will create such a model and compare it with experimental findings on the aforementioned languages and others. I will work closely with linguists and cognitive scientists from differing subfields to ensure that this work benefits from both theoretical and experimental insights and provides a plausible account of acquisition.

Aim 2: To extend this model by incorporating models of other portions of language acquisition. Morphological acquisition does not happen in isolation, and morphology is known to interact with other levels of linguistic representation, particularly phonology (e.g. *-s* is pronounced /s/ in *cats* but /z/ in *dogs*). The nature of the input to the models discussed above assumes that the child has already learned much of the phonology of their native language and extracted the relevant semantic features such as person, number and tense onto which they will map the segmented input. The grouping of forms as discussed in **Aim 1** makes the additional assumption that the child is able to form these groups. To create a more holistic account of acquisition, I will thus integrate models of morphological learning like the above with models of acquisition of other levels of linguistic representation, particularly those on which **Aim 1** relies. It is highly plausible that similar learning algorithms are used for each level of representation, so I will begin by testing the ability of the algorithm developed above to account for these other levels. I will also test integration of the model developed under **Aim 1** with existing models in the literature. With the end goal of an algorithmic hypothesis about how children acquire their native language, I will collaborate with experts on each of the levels of representation I will consider. I will compare the integrated model's predictions with experimental findings on a typologically diverse set of languages to ensure that this algorithmic account of learning is a plausible and generalizable one.

Intellectual Merit. The end goal of this work, an input-to-grammar model of morphological acquisition, will provide insight into linguistic theory and the learning mechanisms employed by children. The model developed under **Aim 1** will provide an algorithmic hypothesis regarding how children learn from skewed, sparse data, and structural hypotheses regarding morphological knowledge in the mind as the end result of acquisition. Both will be valuable in answering questions of learnability and the structure of linguistic knowledge, and may also provide insight into atypical language development. The models developed under **Aim 2** will yield hypotheses about the interactions between linguistic levels of representation, which may be compared with theoretical hypotheses to provide new insight into linguistic structure. Further, these models will give a bottom-up account of language acquisition, and thus yield testable hypotheses regarding the innate factors that may enable it, a highly-debated topic. This model will, to my knowledge, be the first to model morphological acquisition from phonological input to a structured grammar, and it will thus provide a basis for further integrated models of language acquisition.

Broader Impacts. This work has applications to Natural Language Processing (NLP), which focuses on the creation of language technologies. Models used in NLP are typically trained on data several orders of magnitude larger than that to which the child is exposed. This can lead to biases and makes models inaccessible for "low-resource" languages for which large corpora do not exist, such as Indigenous languages and languages of Africa and Asia.[7] Cognitively-motivated approaches already show promising results,[8] and since the algorithms I will develop are designed to succeed on small, sparse input, they will be strong candidates for use with low-resource languages and for testing bias mitigation strategies. This, in turn, will allow for the creation of more accessible, equitable language technologies for all.

References. [1] Brown, R. 1973. *A first language: The early stages*. Harvard University Press. [2] Yang, C. 2016. *The price of linguistic productivity: How children learn to break the rules of language.* MIT Press. [3] Chan, E. 2008. *Structures and distributions in morphological learning*. UPenn Dissertation. [4] Berko, J. 1958. "The child's learning of English morphology." Word. [5] Pinker, S. and Prince, A. 1988. "On language and connectionism: Analysis of a parallel distributed processing model of language acquisition." *Cognition.* [6] Payne, S, et al. 2021. "Learning Morphological Productivity as Meaning-Form Mappings." *Proceedings of the Society for Computation in Linguistics.* [7] Belth, C, Payne S, et al. 2021. "The Greedy and Recursive Search for Morphological Productivity." *Proceedings of the Cognitive Science Society*. [8] Bender, E, et al. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. [9] Xu, Chao et al. 2020. "A Cognitively Motivated Approach to Spatial Information Extraction." *Proceedings of the Third International Workshop on Spatial Language Understanding.*