My research interests lie in using computational models to explore the mental computations and structures that support language learning and processing. In my work, I construct computational models that are informed by theoretical and experimental findings to produce integrated models of the structure of language in the mind. This work has the opportunity to unite perspectives from multiple disciplines—theoretical linguistics, experimental linguistics, and learning theory—by creating explicit models that can be used to verify claims made by each field.

At the University of Pennsylvania, my active participation as a student of the Computer and Information Science Department and an IGERT trainee of the Institute for Research in Cognitive Science has reflected the interdisciplinary nature of my work; I apply the rigorous evaluation of computational modeling required by the computational linguistics community to the problems of language acquisition and processing. As Marr's foundational work expressed, models of cognitive processes can be expressed at several levels—computational, algorithmic, and physical—with each level allowing us to study the problem in a different way. Many recent examples of the application of machine learning techniques to cognitive models have been focused on description at the computational level, expressing *what* the learner is trying to optimize without requiring a cognitively plausible mechanism for *how* and how it could be connected to experimental and theoretical work in linguistics. The currently common practice of modeling at only this level limits the depth of the connection that can be made between models and experimental work. My work aims to supplement and refine work done at the computational level by building models that predict fine details of language behavior, evaluating them on their fidelity to theoretical and experimental findings.

Language acquisition. My interest in modeling was sparked by my early work with Erwin Chan, Mitch Marcus, and Charles Yang on morphology learning. We investigated how a cognitively-motivated morphology learning system could be adapted as a model of child language acquisition. While problems of irregular morphology (e.g., the English past tense debate) had received much attention, no model had yet explained the robust order of acquisition of regular morphology observed by Brown 1973. We found that a greedy approach to learning regular morphological processes that relied on counting the unique words that participate in the process can predict the order of affix acquisition observed in young children.

One problem addressed in my dissertation is word segmentation, the problem of how infants learn to segment continuous speech into words. Computational-level models of word segmentation (e.g., Goldwater et al., 2009) have provided broad insight into the types of cues that infants may attend to in order to identify word boundaries and the way in which they might balance the quality of the lexicon with its expressiveness. However, in order to meaningfully model the development of infants' competence over time and incorporate details of their capabilities identified by experiments, I believe it is necessary to develop online-learning, cognitively-plausible models. The system I have built for simulating infant word segmentation over time is the first to predict the observations of child development researchers that infants first choose over-sized units for words (e.g., *that's-a*) (Brown, 1973) but later overcorrect by selecting under-sized units (e.g., segmenting off *be* in *behave*) (Peters, 1983). This work has been presented in computer science (CoNLL), linguistics (WCCFL), and psychology (BUCLD) venues, affirming the broad appeal of building an integrative, interdisciplinary model.

Language processing. In addition to building models that bring together developmental and experimental perspectives in acquisition, my work seeks to extend experimental findings in language processing by using modeling to explore effects on large-scale data. As part of my dissertation, I am exploring the long-standing impasse that has formed between advocates of dual-route (e.g. Baayen et al., 1997) and decompositional models (e.g. Taft, 2004) of lexical processing. The impact of this debate extends beyond processing; whether whole words are stored for forms that could be derived compositionally bears on the selection of theories of word formation that do (Goldberg, 2006) or do not (Halle and Marantz, 1993) predict such storage. By evaluating models using the English Lexicon Project (ELP), the largest set of visual lexical decision times collected (Balota et al., 2007), and applying best practices in mixed-effects modeling, I have found that the data are consistent with decompositional approaches and demonstrated that the most compelling evidence for the dual-route model given in Alegre and Gordon 1999 can be traced to a combination of poor frequency estimates and poor statistical practices. The identification of the computational properties of a decompositional model has enabled the development of an algorithmic-level model of morphological processing.

Future research goals. I have a number of additional research questions that I am eager to explore. I have thus far analyzed reaction times to words in isolation in a visual task. I look forward to expanding this work into sentence processing and evaluating neural measures of morphological decomposition in addition to behavioral ones. While the ELP has provided a crucial starting point for large-scale evaluation of models of morphological processing, I plan to create a web-deployable framework to collect an auditory lexical decision data set of similar scale to further our understanding of lexical access in more natural settings and phonological influences on processing.

I also look forward to expanding the methods of evaluating models of language acquisition. Well-recorded developmental patterns such as age of acquisition can be used as another reference for evaluating the quality of word segmentation models. For example, the words that are acquired early during development—as recorded in developmental data such as CDI norms—should also be reliably identified at early stages of learning in simulation. More broadly, I intend to extend the holistic approach I have taken to modeling development to also compare the predictions of models with corpora of child productions over time.

In sum, my research provides a testing ground for models of language in the mind and reaps the benefit of integrating theory, experimental work, and validation against large-scale data sets.

References

Alegre, M. and Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40(1):41–61.

Baayen, R., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1):94–117.

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445.

Brown, R. (1973). *A First Language: The Early Stages*. Harvard University Press, Cambridge, Massachusetts.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language.* Oxford University Press, Oxford.

Goldwater, S., Griffiths, T., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Halle, M. and Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K. and Keyser, S., editors, *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*, pages 111–176. MIT Press, Cambridge.

Peters, A. (1983). *The Units of Language Acquisition*. Cambridge University Press.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57A(4):745–765.