New Tools for Understanding Language Learning and Generalization

Perhaps the most celebrated property of natural language is creativity, the ability to combine existing units to derive new expressions. This feature is found across multiple levels of representation: Pre-existing phones can be combined to form novel morphemes, which can be combined to form novel words, which can be combined to form novel sentences. But linguistic creativity is also bounded, giving rise to a number of long-standing puzzles of generalization. A syllable like *derp* sounds like a better morpheme of English than a syllable like *denp*. The words *warmth* and *truth* exist in English, but *coolth* does not. And speakers of English will naturally drop the phrase *on the counter* from *John made dinner on the counter*, but not from *John put dinner on the counter*.

My work aims to explain how languages strike a balance between productive generalization, on one hand, and conservative reuse of attested patterns, on the other. The core idea is that language learners acquire a system of rules and stored items, which optimizes a *tradeoff* between a pressure to explain the input with the fewest, most general rules possible, and a countervailing pressure to predict specific, idiosyncratic properties of the data. This idea is old, lying at the heart of many theories of optimal inductive inference. However, recent advances in probabilistic and statistical computation have made it possible to exploit the full power of this tradeoff-based approach and apply such *idealized learning* frameworks to complex, highly-structured models and very large datasets. These tools are especially well-suited to the hierarchical, recursive, and compositional structures of natural language, making it possible to derive new kinds of predictions from linguistic models and to connect theoretical assumptions with new kinds of evidence.

In my research, I study problems of language learning and generalization by combining theories of representation drawn from the linguistics literature with modern tools of statistical computation. Below, I describe several case studies which use Bayesian methods to derive idealized learning predictions for formalized and implemented linguistic proposals. These studies illustrate the multiple ways in which this approach can contribute new insights to our understanding of linguistic creativity. In some cases, idealized learners can provide first principle explanations for phenomena that have required stipulation in other theories. In other cases, idealized learning models can provide evidence for or against different linguistic representations or mechanisms. Idealized learning models can also help us understand the complex ways in which linguistic modules can interact over the course of learning. Finally, by combining statistical methods and linguistic theory, it is often possible to derive new kinds of theoretical predictors and to connect theoretical predictions to new kinds of empirical dataat previously impossible scales.

Learning the Pattern of Morphological Productivity One domain which in which the tension between productive generalization and conservative reuse has been a long-standing puzzle is morphology. How do learners determine that of the three semantically similar suffixes, -*ness*, -*ity*, and -*th*, -*ness* generalizes freely (e.g., *Lady-Gagaesqueness*, *pine-scentedness*), -*ity* generalizes in certain morphological contexts, but not others (e.g., after -*able* in *tweetability*, but not after -*less* in **cluelessity*), while -*th* never generalizes—despite appearing in many existing words (e.g., **coolth* despite *truth*, *width*, *warmth*).

In a recent MIT Press monograph, I proposed a new theory to address this question [1].[†] The model is based on the idea that when encountering a morphologically complex word like *truthiness*, the learner considers all of the possible combinations of productive competition and storage which could have given rise to the word: [*truthiness*], [*truth*]+[-*y*]+[-*ness*], [*truthy*]+[*ness*], [*truth*]+[-*iness*]. By storing complex, whole forms (e.g., [*truthiness*]) individual derivations will be simple—consisting of a single retrieval step—but the lexicon will be large. By using a greater amount of decomposition (e.g., [*truth*]+[-*y*]+[-*ness*]), the

[†] Numbers refer to entries in curriculum vitae.

size of the lexicon can be minimized, at the expense of a greater number of computational steps per word. The model uses Bayesian inference to choose the derivation of each form it observes, optimizing a global tradeoff between simpler derivations and fewer, simpler lexical items.

To evaluate this *inference-based* approach to productivity, I conducted an idealized learner study, comparing this proposal with three other approaches from the literature. In the *full-listing* model, structures are composed the first time they are encountered, but stored and reused as wholes thereafter. This proposal captures the classical notion of lexical redundancy rules from generative morphology. In the *full-parsing* model, all words are fully decomposed, providing a baseline for the other models. Finally, the *exemplar-based* approach attempts to store every generalization consistent with the data, hedging across multiple different hypotheses. In the book, I present several alternative formalizations of these ideas, and connect them to several core ideas from the theory of programming languages.

I test these models' ability to account for the correct pattern of generalization for a variety of phenomena from English inflectional and derivational morphology. For inflectional morphology, only the inference-based model is able to learn the correct pattern of *defaultness* of the regular +/d/ rule, together with the phenomena of *blocking* of the regular rule by the existence of irregular forms like *went*. In order to explain blocking, most proposals in the linguistics literature have used a variant of the *elsewhere condition*, which states that a rule with more specific input conditions takes precedence over a rule with more general conditions. I show how a variant of the elsewhere condition follows as a necessary consequence of the assumptions of the inference-based model, without the need for further stipulation. I also show how the inference-based model explains a number of reaction time phenomena from the psychology literature as well as overregularization phenomena from development.

Turning to English derivational morphology, I show that only the inference-based model makes plausible predictions about the productivity of English derivational morphemes like -*ness* and -*ion*. I also show how the leading quantitative theory of productivity, Baayen's *hapax-based* approach, can be understood as a special case of the inference-based model. The model also explains a number of well-known phenomena of affix ordering. For example, in words containing more than one affix, more productive affixes tend to appear outside of less productive affixes. The inference-based model explains this generalization, but also predicts a number of exceptions unaccounted for by earlier theories—-such as generalizable suffix combinations like -*ability* and -*ation*.

In other work, I have shown how the productivity model can explain other kinds of data, such as the acquisition pattern of dative verbs [13], reading time latencies in eye-tracking corpora [3], and word-frequency-estimation errors [18,39]. One interesting and unexpected prediction of the model is that the existence of high-frequency irregular forms, such as *went*, can make the productivity of regular rules, such as +/d/, easier to learn. Together with Kenny Smith, I have recently provided experimental evidence in an artificial language learning setting for such an irregularization bias, demonstrating that high-frequency irregulars facilitate the generalization of regular rules for human language learners, and conversely that high-frequency regulars inhibit the generalization of a regular rule—a result with implications for language typology and change [23,30].

Learning Phonotactic Generalizations Another mechanism for linguistic creativity is the generation of new morphemes via processes such as borrowing (e.g., *sudoku*), blending (e.g., *spork*), pronunciation of acronyms (e.g., *laser*), and invention from the whole cloth (e.g., *derp*). The adoption of such morphemes is governed by systematic intuitions about which sequences of sounds are likely and unlikely in a language.

With Richard Futrell and Adam Albright, I have developed a novel approach to learning such *phonotactic generalizations* [19]. This model differs from earlier approaches in the way that it conceptualizes phonotactic learning. Most earlier models represent phonotactic generalizations in terms of constraints that penalize illicit combinations of phonological features, such as *-*np*. By contrast, our model treats the problem as one of learning a *phonotactic lexicon* of reusable sub-morphemic units such as -*rp*. Under this approach, a sequence like *derp* is a likely morpheme of English because it reuses attested phonotactic patterns, rather than because its fails to violate phonotactic constraints.

Our model is built on several insights from phonological theory. Like most models of phonology, we start by assuming that phones are represented in terms of *phonological features* which capture shared articulatory and acoustic properties of sounds, like the fact that the phones /n/ and /m/ are both articulated with a lowered velum (i.e., NASALITY). To accurately capture phonotactic generalizations, we make use of three ideas about the representation and organization of phonological features.

First, inspired by work on *feature geometries* and the *contrastive hierarchy*, we assume that features can be organized into a cross-linguistically universal *feature dependency graph*. This graph encodes universal contingencies between features such as the fact that only consonants, but not vowels, specify LATERALITY, while both consonants and vowels specify NASALITY. Second, inspired by the idea of *autosegments*, we assume that the phonotactic lexicon stores not only patterns involving fully-specified phones (e.g., -*rp*), but also patterns involving sub-phonemic bundles of features, for example, patterns involving vowels whose nasality is unspecified. Third, inspired by the idea of phonological tiers, we assume that the structure-building operations which combine stored phonotactic patterns are sensitive not only to immediately adjacent units, but also to distal units which occupy similar parts of phonological space, allowing the model to capture phenomena such as the tendency for vowels to share the same feature values within words (i.e., *vowel harmony*).

The model performs an inference to determine which reusable phonotactic patterns, at varying levels of abstraction, best characterize the input data. For example, if the input contains input words like *twerp* and *burp*, the model determines whether the language is better characterized by a pattern allowing the specific phone sequence -*rp*, or by more abstract patterns, such as an approximant followed by a stop. To perform this inference, the model optimizes a tradeoff between simpler but more permissive stored patterns (e.g. consonant-consonant) which can weakly characterize a large number of inputs, and more specific patterns (e.g., -*rp*) which more tightly characterize a smaller number of inputs.

We performed an idealized learner analysis that systematically compared variant models which did or did not include each of these three ideas about feature organization. Our results indicate that each of the three representational devices consistently improves the model's ability to predict the phonotactic well-formedness of novel words from a sample of fourteen typologically diverse languages. In order to better understand how and why the models behave as they do, we developed several novel information-theoretic analyses of model behavior. For example, we used a measure of representativeness to provide example word forms which are most likely to discriminate between models. In another analysis, we showed that differences in model behavior were driven in large part by MANNER-feature interactions.

Learning the Argument Structure of Lexical Items The ways in which more complex expressions can be built from lexical units are highly constrained by the *argument structure* requirements of those lexical items. For example, when used in a sentence such as *John put the loaf of bread on the counter*, the verb *put* must appear alongside (i) a subject noun phrase expressing which is doing the putting, (ii) an object noun phrase expressing what is being put, and (iii) a prepositional phrase expressing the destination of the putting event.

Since much grammatical structure must be explained by lexically-specified argument structure requirements, it is natural to ask whether most or all structure is lexically-specified or whether the grammar must possess non-argument modes of composition. One important empirical phenomenon relevant to this question is the existence of *modifier* phrases that do not obviously satisfy argument-structure requirements—such as the adverbial phrases *thoughtlessly* and *while preparing dinner* in the sentence *While preparing dinner, John thoughtlessly put the loaf of bread on the counter*. While most grammatical theories have historically proposed extra-lexical mechanisms for composing such phrases (usually formalized as a form of *adjunction*), there has been a great deal of debate about which phrases are modifiers, what empirical phenomena are relevant to this question, and the machinery needed to handle these phenomena. Given the contention in the literature, it is important to ask whether new kinds of evidence can be brought to bear on the debate.

With Leon Bergen and Ted Gibson, I have conducted a study examining the question of whether the distribution of constituents in natural language can be more parsimoniously explained by a grammatical system that includes some non-argument mode(s) of composition, or whether an argument-structure-only model suffices to explain distributional aspects of the linguistic data [10,22]. We identified three ways in which arguments and modifiers are predicted to differ in distribution across nearly all theoretical proposals. First, lexical items specify a finite number of arguments, while an unbounded number of modifiers can be composed with constituents. Second, arguments are typically obligatory, while modifiers are always optional. Third, arguments tend to stand in fixed structural relations with their selecting lexical item, while modifiers exhibit a greater degree of structural flexibility.

We performed an idealized learner study comparing the predictions of an argument-only model with an argument-modifier extension of that model, where both models were designed to minimally capture the three core distributional differences above. As in the preceding projects, learning for each model was governed by a tradeoff between pressure for a smaller, simpler lexicon, on one hand, and pressure for simpler sentential derivations on the other.

Our results indicate that there is clear distributional evidence for an argument-modifier distinction with the three distributional consequences above and that the argument-modifier status of many constituents is learnable from these distributional cues. The argument-modifier model was able to recover the argument-modifier status of many individual constituents when evaluated against a hand-annotated gold standard. Moreover, it (i) provides a simpler account of the input data both in terms of the size of the lexicon and the complexity of derivations of individual sentences and (ii) generalizes more robustly to novel sentences. Intuitively, by identifying modifier phrases, the argument-modifier model is able reuse lexical argument structures across a greater variety of sentential constructions, for instance, using the same argument structure for the verb *put* across the sentences *John thoughtlessly put the loaf of bread on the counter* and *John put the loaf of bread on the counter*. This work demonstrates how idealized learner theories can provide new kinds of evidence about representational and mechanistic questions in linguistics, and can serve as a template for the application of these techniques to other problems.

Learning a Lexicon from Acoustic Input Each of the preceding projects can be seen as studying one component in isolation of the overall problem of learning the lexicon of a language. In another strand of work, with Jackie Lee and Jim Glass, I have adopted a different approach, building a joint learning model which attempts to solve multiple parts of the problem in parallel [4].

Our system consists of three components. First, at the bottom layer, the model acquires a set of units which characterize the acoustic properties of phones and segment the input stream. Second, the model acquires lexicons of subword- and word-like units, consisting of hierarchically nested sequences of phones. Between these two components is a *transduction* module which learns to map the phone-like units that constitute underlying lexical items to the surface realizations of those phones. This module handles symbolic variation arising from phonological and phonetic processes such as allophony and coarticulation. All of these components are unsupervised, and do not assume anything in advance other than the existence of units like phones, morphemes, and words, and a distinction between underlying and surface realization of phones.

In our study, we perform an idealized learner analysis, systematically comparing variants of the model with and without the lexical, acoustic, and transduction components of the model. The model achieves performance that is competitive with state-of-the-art spoken term discovery systems. More importantly, our results provide preliminary evidence that simultaneously learning sound and lexical structure can lead to synergistic interactions whereby the full model outperforms variants that do not make use of the transduction or acoustic components. We also performed a number of detailed analyses of the kinds of linguistic structures the model learned, with a number of interesting results. For example, the model stores a frequently-reused lexical unit corresponding to the suffix sequence *-ation*—a suffix combination predicted to be very generalizable by the productivity model discussed above. The fact that this generalization can be recovered in an entirely unsupervised fashion from acoustic data provides further converging evidence in favor of the tradeoff-based approach to lexicon learning.

Ongoing Work and Future Directions I have a number of ongoing projects that apply the idealized learner approach to a range of other linguistic phenomena. With Leon Bergen and Roni Katzir, I am studying the learnability of various mechanisms for handling *wh-* and other kinds of movement. With Michelle Fullwood, I have developed a model of lexicon learning for non-concatenative morphological systems, such as the Arabic verbal system [10]. We are extending this approach to capture ideas from the prosodic morphology literature. With Kevin Ellis, I am developing models of phonological rule learning, based on the paradigm *Bayesian program learning*, which generalizes a number of earlier approaches in the literature. We are currently examining the ability of this model to capture vowel harmony and floating tone systems. I have a long-standing interest in lexical semantics and verb-argument structure [6,13,14]. With Josh Hartshorne, I am conducting an idealized learner study examining the feasibility of different theories of verb meaning to explain the distribution and learning of verb-argument constructions.

Practical implementations of the kinds of models described above often require grappling with fundamental issues in the design of inference algorithms. I have been exploring a number of general techniques for providing faster, more robust inference for linguistic models. With Tejas Kulkarni, I am developing hybrid generative-discriminative architectures for parsing. These pair powerful-but-slow, top-down generative models with brittle-but-fast, bottom-up neural network processing. In another line of work, I have been exploring the idea of *self-relaxed* parsing algorithms which add systematic, parameterized noise to discrete parsing problems in order to smooth the search space. Finally, in a number of ongoing projects, I am building new resources for testing models of linguistic creativity, including compiling large databases of crowd-sourced judgments about suffix generalizability, to test models of productivity, and wordlikeness, to test models of phonotactics.

My future research will focus on building integrated learning models that more accurately capture linguistic assumptions and cover a wider variety of empirical phenomena. For example, within the next few years it will likely be possible to build a model of acoustic lexicon learning like [4] that more accurately captures productivity and phonotactics generalizations [1,19], non-concatenative structure-building [10], and argument-structure restrictions [9.20]. Building and testing such integrated models of learning will require developing new inference algorithms, new methods for comparing and falsifying models, new data sets, and new forms of argumentation. I look forward to contributing to solutions to all of these problems.