Research Statement

Why is language the way it is? I'm interested in questions like: Why are the relative orders of object and verb correlated with the relative orders of preposition and noun? Why do some languages use word order to signal predicate-argument structure, while others don't? Why do some languages have grammatical gender systems? At a more basic level, why do all languages have words, ambiguity, and syntax that deviates only slightly from hierarchical embedding?

My research tries to go as far as possible with the hypothesis that the answer to these questions lies in understanding language as an efficient communication system for agents with human-like information processing constraints. That is, languages should evolve to enable communication while minimizing processing cost for the average utterance. To address this hypothesis, I develop computational models of human language processing and test how well they predict the quantitative distribution of structures in multiple languages. I see this work as complementary to the traditional approach in formal linguistics, where innate representational constraints give rise to universals.

In computational psycholinguistics, my goal is to articulate and test models of human communication and information processing which are accurate and yet simple enough to derive quantitative predictions about typology and production preferences. One long-standing theory along these lines is dependency length minimization (Hawkins, 1994); I have formulated and validated the quantitative predictions of this theory in parsed corpora of dozens of languages, finding the predictions hold in all languages studied (Futrell, Mahowald & Gibson, 2015; PNAS). On the theoretical side, I have used the highly general noisy-channel theory of sentence processing to develop a model of processing difficulty that derives and generalizes dependency locality effects, yielding a rich set of new psycholinguistic and typological predictions (Futrell & Levy, 2017; EACL).

This research program also requires quantitative characterization of linguistic structure in usage, as it might be observed in a corpus. I have worked to link two apparently orthogonal levels of description, typically studied by disjoint research communities: language as a productive formal system that expresses meanings, and language as a probability distribution over morphemes and sentences observable in usage. To this end, I studied methods for quantifying word order freedom in linguistic and statistical depth (Futrell, Mahowald & Gibson, 2016; DepLing), as well as the statistical correlates of autosegmental structure in phonotactics (Futrell, Albright, Graff, & O'Donnell, 2017; TACL).

My knowledge and experience make me well-positioned to carry out this research program. As an undergraduate and Master's student at Stanford, I studied syntax, corpus linguistics, and NLP. I have been carrying out psycholinguistic experiments to test processing theories since my undergraduate years. I have industry experience building natural language understanding systems. As a PhD student at MIT, I gained a firm understanding of probabilistic models, modern cognitive science, and computer science.

Dependency Length Minimization

One long-standing theory of how processing efficiency gives rise to linguistic universals is dependency length minimization (DLM): the idea that production preferences and typological distributions can be explained by a pressure to keep the distance between words in a syntactic dependency short. This pressure is motivated by **dependency locality effects** in sentence processing: processing difficulty seems to increase when syntactically related words in a sentence are distant, for reasons related to working memory limitations. DLM has been offered as a high-level explanation for the Greenbergian harmonic word order correlations (Hawkins, 1991) as well as for fundamental properties of grammar such as the overwhelming frequency of projective or context-free structures (Ferrer i Cancho, 2006).

I have provided substantial empirical support for DLM as a theory of word order patterns. In Futrell, Mahowald & Gibson (2015; PNAS), using recently-available crosslinguistic parsed treebanks, we address the issue of whether corpora of 37 languages show evidence for DLM beyond what one would expect from independent constraints for context-freeness, fixed word order, and consistent head direction; these constraints have independent motivations based on learnability and parsing complexity. Without exception among these languages, we find that attested dependency length is shorter than what one would expect from the independent constraints. The result constitutes the largest-scale evidence for DLM as an explanatory principle for word order in grammar or usage across multiple languages.

This work also raises two new questions. First, the result leaves it unclear whether DLM in corpora is better explained by grammar or usage preferences. For instance, English grammar licenses the orders (1) and (2); (2) has a long dependency between *threw* and *out*.

(1) Kim threw out the old trash that had been sitting in the kitchen for several days.

(2) Kim threw the old trash that had been sitting in the kitchen for several days out.

The observed DLM effect may be due to speakers choosing orders with low dependency length such as (1) from the set of grammatical orders, without the grammar itself being affected by DLM. To determine whether DLM affects both usage and grammar, I developed a probabilistic model of the grammatical orders of a dependency tree (described in Futrell & Gibson, 2015; EMNLP), so the attested orders can be compared to counterfactual grammatical orders. If real orders are shorter than grammatical orders, then we have evidence for DLM in usage. And if random grammatical orders are shorter than random baseline orders, then we have evidence that DLM affects grammar, here construed as the set of licit orders for a dependency tree. These predictions come out, suggesting DLM affects both usage and grammar.

Second, the large-scale corpus approach reveals that there is residual variance between languages that is not explained by DLM alone. I have found that DLM is weaker in languages with more head-final structures such as Japanese than in languages with more head-initial structures such as Indonesian. Applying measures of word order freedom discussed below, I also found that languages with more word order freedom have less DLM pressure. These results constitute explananda requiring a theory beyond simple DLM, which has motivated further theoretical work.

Information Locality

DLM as a functional theory relies on the notion of locality effects in sentence processing. Yet the major effects on processing difficulty seem to come from surprisal: the probability of a word given its previous context. In recent work, I have focused on developing a simple unified model of processing difficulty that integrates surprisal with locality effects, simultaneously solving some outstanding puzzles in psycholinguistics and yielding a simple generalization about processing difficulty and production preferences which we call information locality.

In Futrell & Levy (2017; EACL), I develop a model where the processing cost of a word is the surprisal of the word given a *noisy* representation of the preceding context, where the comprehender corrects for noise in the representation by doing noisy-channel inference. I use the model to give the first formal explanation for a long-standing puzzle in sentence processing: language-dependent structural forgetting effects, which create cases where an ungrammatical sentence such as (3) seems more acceptable than a grammatical one such as (4) in English, while the preferences are reversed for equivalent sentences in German (Gibson & Thomas, 1999; Vasishth et al., 2010; Frank et al., 2016).

(3) *The apartment that the maid who the cleaning service sent over was well-decorated.

(4) The apartment that the maid who the cleaning service sent over cleaned was well-decorated.

In this model, a series of verb-final relative clauses is unlikely to be maintained correctly in a noisy memory representation when such clauses are rare (as in English), resulting in harder processing for the grammatical sentence. But these clauses are easily maintained in memory when they are common (as in German). The model shows precisely how language statistics, resulting from grammatical differences, can drive language-dependent memory effects in processing.

I also show that the model gives rise to a new generalization, **information locality**: that processing is easiest when words that are highly associated with each other (have high mutual information) are near to each other. For example, the words "give" and "up" in English are highly associated in that they mutually predict each other, appearing commonly together as a phrasal verb. I show that words in syntactic dependencies have relatively high association, meaning that dependency locality effects can be subsumed under information locality effects. This work constitutes the first explanation of dependency locality effects in a high-level expectation-based framework.

Quantitative Analysis of Syntax and Other Linguistic Structure

In my view, the fundamental problem for quantitative studies of syntax is to link two levels of linguistic description. The first notion is the static syntactic structure of a sentence, the network of relations between words which define how the utterance composes to get its meaning or how the

individual sentence was abstractly generated. The second notion is distributional: characterizing the statistical co-occurrence patterns among words and phrases as they can be observed in usage.

Approaches emphasizing one level of description have tended to take a reductive view of the other. For instance, Stefan Frank's group has argued that language processing needs no notion of hierarchical structure, and co-occurrence statistics suffice to explain language understanding. On the other hand, Chomsky argued that probabilities are irrelevant to syntax: for instance that the frequency of "New York" after "to" has no relevance for our understanding of prepositional phrases. I believe these mutually reductive views are misguided: both levels of description are necessary for understanding processing and learning. Yet a rigorous understanding of how the two levels relate has remained elusive.

Understanding the connection between statistical and syntactic structure requires large parsed crosslinguistic datasets which have only recently become available. I have been at the forefront of the effort to use and evaluate recent datasets for quantitative linguistic studies; in particular, my work is among the first to use corpora of the Universal Dependencies project (UD; Nivre et al., 2015) for studying crosslinguistic quantitative phenomena such as word order freedom, dependency length minimization, and the statistical correlates of dependency structure, and to strengthen the foundation for such studies by evaluating the strengths and weaknesses of these corpora.

For example, in studying word order freedom (Futrell, Mahowald & Gibson, 2015; DepLing), I showed how to reconcile a theoretical best measure of the degree of word order freedom with what is feasible to measure from UD corpora. I discuss the validity of the resulting measure in linguistic, statistical, and practical detail, taking into account the annotation standards of the UD corpora. In the end, I find that intelligible and robust estimates are indeed possible for a subset of cases of word-order freedom. The measures support a model of word order variation advocated in Gibson et al. (2013), with development in Futrell et al. (2015; Cognition), holding that SOV order implies morphological marking of verbal arguments even when word order is fixed, for reasons of communicative robustness.

In addition to work using UD corpora, I have worked to produce a freely-available parsed corpus of Pirahã, a language at the center of bitter controversy about whether it has recursive embedding. Taking a careful corpus-based approach to the question, and comparing to English corpora, we conclude that the Pirahã corpus gives no unambiguous evidence for embedding, but also no strong evidence against it (Futrell et al., 2016; PLOS ONE).

While my main interest in quantitative structure of language is on syntax, my interests are broad and in particular I have produced a probabilistic interpretation of autosegmental structure in phonotactics (Futrell, Albright, Graff & O'Donnell, 2017; TACL). Implicit in the model is a fundamental postulate about how autosegmental structure relates to statistical structure: that feature groupings which are more closely bound together in a feature geometry will co-occur with a distribution very different from what one would expect from the chance co-occurrence. This linking hypothesis between statistical distributions and linguistic structures is reminiscent of the connection between syntactic and statistical dependency discussed above under Information Locality.

Future Work

The principle of information locality—that words which predict each other should be close—can provide rich predictions and explanations for syntax, touching on the relative orders of adjectives, arguments and adjuncts, and the effects of agreement morphology, which increases mutual predictability among words. Diachronically, information locality can be seen as attraction between statistically associated words, giving a mathematical theory of word order change and grammaticalization paths. These are some avenues of future work.

The time is ripe to ask questions at the intersection of computational processing theories and linguistic theory. As faculty I plan to set my sights on broader set of linguistic universals, both those known from classical typology and potentially novel quantitative ones. For example, why do animate NPs typically appear before inanimate ones, as both a soft and hard constraint? Why are binding domains usually restricted to the clause across languages? Why do head-final languages typically have dependent-marking morphology, but head-initial languages usually head-marking? I have formulated tentative processing efficiency hypotheses for these problems, and look forward to to leading projects to develop and test those explanations as an advisor.