JON A. WILLITS

RESEARCH STATEMENT

How do humans learn language? Children acquire this knowledge so quickly, and we sometimes forget how incredible this is. The goal of my research is to better understand the mechanisms underlying the mastery of linguistic knowledge. Recent advances in our ability to study large, naturalistic datasets, combined with advanced computational modeling techniques, have allowed us to ask questions that were not possible just years before. One of the major insights from my work is that "Big Data" research that investigates the structure of experience – such as investigations of giant corpora of naturalistic speech – will force us to radically re-evaluate certain theories of learning and representation. Many models (especially any that involve interactions or nonlinear effects) perform qualitatively differently when faced with large amounts of data, making some learning problems harder and others *much* easier.

Programmatically, My research is strongly interdisciplinary, and sits at the junction of cognitive psychology and a number of other disciplines. Much of my work takes a developmental perspective, and theories and methods from linguistics play an important role. Because much of my research is computational, my background (and B.Sc. degree) in computer science and machine learning are extremely important. One of my goals is to model the development of knowledge within neurologically plausible systems, and ideas from developmental neuroscience inform my work. Finally, my approach involves overlap with other domains, resulting in a number of projects and collaborations in social cognition, clinical psychology, and speech and language disorders.

Theoretically, I take a developmental and interactionist perspective. My research program is driven by three guiding principles. First, the cognitive system is driven by highly complex and structured environmental input. Second, input is processed by perceptual and learning systems that are constrained and biased, privileging certain types of outcome representations. Third, the learning process is best understood in terms of its trajectory over time. The mind is a dynamical system – what you knew before strongly influences how you perceive the world, and what you are likely to learn from new experiences.

Methodologically, my work has three sub-components: (1) analyses of large, naturalistic datasets, (2) computational models of learning and knowledge use, and (3) behavioral experiments with infants, older children, and adults that test model predictions. "Big Data" analyses, when used in theory-driven ways, can lead to new explanations of the development of complex behaviors. These explanations are then tested empirically, and the results used to help refine the models.

Topically, I study a number of questions related to the acquisition of language, and how the structure of our experiences can cause changes in knowledge and behavior. Most of this research demonstrates how the constrained learning of complex environmental input leads to structured knowledge, and how a major component of this learning process involves using prior knowledge to learn about new experiences. Below I describe a few examples in detail.

How do "Big Data" and Computational Capabilities Interact During Learning?

In order to understand the nature of the computational system that people use to learn and represent knowledge, we need to document the quantity and quality of the input that they receive. The interaction between the input and the computational system is complex, and sometimes our intuitions about the input, and what certain computational systems will do with that input, are mistaken.

One example is my dissertation research (Willits, 2012; Willits, 2013), in which I modeled the acquisition of various complex linguistic structures. I showed that a single learning model, the simple recurrent neural network, can be used to simulate the acquisition of a wide range of different linguistic phenomena. These results suggest that statistical learning models, including recurrent neural networks, need not be seen as antagonistic to structured explanations of knowledge. Instead, they may play a role in helping us understand the mechanisms by which those structures are learned.

Another example concerns the nature of children's semantic representations. By three years of age,

children appear to be biased towards taxonomic, hierarchically structured representations of semantic knowledge. This has led some to hypothesize that the cognitive system is hard-wired to prefer these kinds of representations. In my research, I have shown that the structure of speech to children (using corpora of millions of words of child-directed speech) is *itself* extremely structured and organized (see **Fig. 1**, for an example). This suggests that a relatively small degree of biasing of the system is needed in order to result in hierarchically structured representations (Willits & Jones, in review).

I have also collaborated on projects comparing the capabilities of various models of learning of semantic structure (Rubin et al., 2014; Willits, Rubin, & Jones, in review), and reviewing these models to situate them within broader theoretical frameworks (Dennis, Jones, & Willits, 2014; Willits, in preparation). Finally, testing models of language development is difficult, but critical. To this end, I developed a novel method for testing semantic priming in infants using the Headturn Preference Procedure (Willits, Wojcik, Seidenberg, & Saffran, 2013), which will be helpful for understanding more about children's semantic knowledge and its structure at extremely young ages.

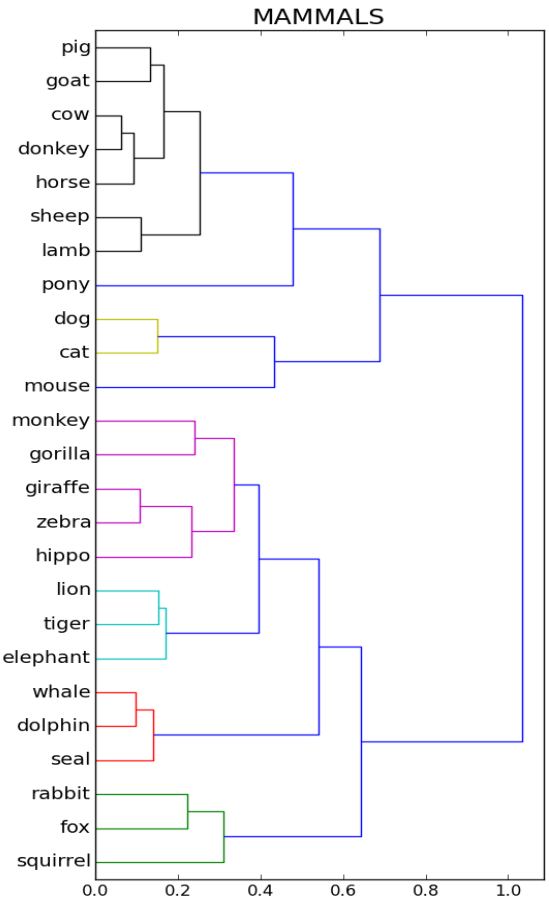

Fig 1. A hierarchical structure diagram of "mammals" produced by a model that discovers semantic organization from 6 million words of child-directed speech (Willits & Jones, in review).

How Do Meaning and Structure Interact?

Psychological research often compartmentalizes different domains of knowledge (for example, semantic versus syntactic knowledge). In my work, I try to draw attention to the potential learning boosts that come from *integrating* knowledge across domains. In many domains the learning problem is often characterized as intractable without strong innate constraints. In some cases this may be true, but in many cases these learning problems are made artificially more difficult by assuming that domains are learned independently of one another. One example involves children's learning of complex syntactic structure. My research suggests that a major difference between easy-to-learn structures and hard-to-learn structures is that difficult structures lack external cues, (such as correlated perceptual or semantic information), which draw attention to the structure to be learned. When correlated cues are present (see **Fig 2**. for an example), learning is facilitated (Willits, Lany, & Saffran, in review, Willits & Saffran, in review, Willits, in preparation). By treating different domains of knowledge as distinct, we risk missing insights about their acquisition. This research is an example of the ways in which prior knowledge scaffolds future learning, and how the learning process is best understood as a complex interaction that unfolds over time.

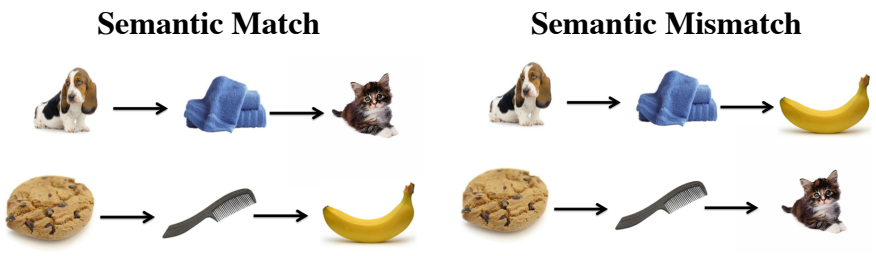

Fig 2. Toddlers learn long-distance dependencies – such as that item 1 will perfectly predict item 3 – when those items share a semantic cue (e.g. both are animals or foods, as shown in the left panel), but not when they don't (as shown in the right panel; Willits, Lany, & Saffran, in review).

The assumption that different domains of knowledge are independent leads us astray in other domains as well. Word meaning, for example, involves a complex interaction of information from both linguistic and nonlinguistic sources. In many cases, word meanings are highly grounded and perceptual. But people typically use language to highlight unusual situations, not to talk about obvious, common things ("*purple carrot*", for example, is more frequently spoken than "*orange carrot*"). In a number of experiments, I have documented that people are sensitive to the differences between grounded semantic information and semantic information acquired from language (Willits, Amato, & MacDonald, 2015; Willits et al., 2007; Willits & Seidenberg 2008; Willits & Jones, in review). Linguistic information and grounded information come to us by different mechanisms, but a proper understanding of word meaning will entail explaining how those types of knowledge interact during learning and use.

What Are the "Units" in Statistical Learning?

Another example of prior knowledge impacting future learning concerns the units over which learners are tracking information. We know that children rapidly learn statistical dependencies in their environment. In order to formalize theories of statistical learning, we need to establish how children come to identify the units over which they are tracking statistics. Because once certain units become established, they bias future learning, due to those units (and not others) being the objects of attention. Understanding this phenomenon will help explain many puzzles in acquisition.

One such puzzle is why verbs are harder to learn than nouns, a fact that has its genesis in infancy. Previous research has shown that by 7 months, infants recognize nouns in speech, but that they don't recognize verbs until 4-6 months later. I performed corpus analyses showing that the statistics of English *should* promote the recognition of nouns relative to verbs; nouns occur in more statistically "useful" distributions (Willits, Seidenberg, & Saffran, 2009). However, this noun advantage only holds for traditionally defined words. If one instead assumes infants are tracking statistics at other levels (i.e., treating chunks such as *–ing* as separate units), the calculus changes dramatically (see **Fig 3**). In the presence of *–ing*, verbs ought to be just as easy to recognize as nouns (because, among other reasons, *-ing* can serve as a useful anchor cue, in a manner similar to how words like *the* do so for nouns).

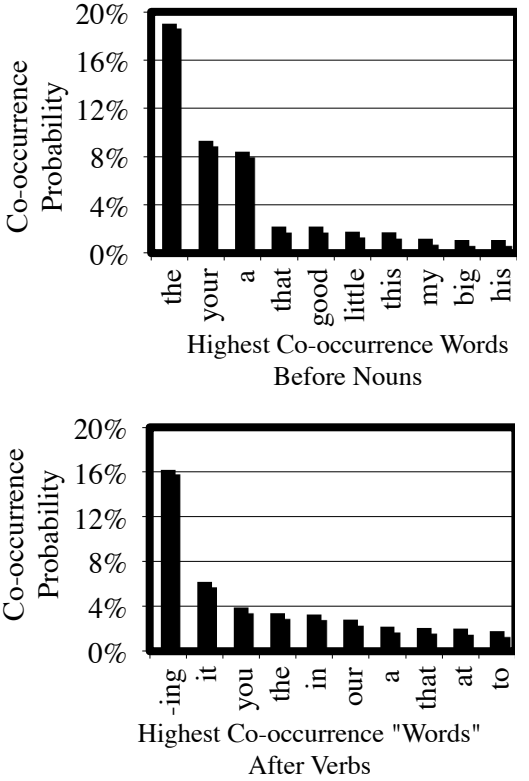

Fig 3. Co-occurrence probabilities for the most-frequent nouns/verbs in child-directed speech (Willits, Seidenberg, & Saffran, 2014).

We then tested these predictions in word recognition experiments with 7-month-old infants, and the results demonstrated that while infants have trouble recognizing familiar verbs without the *–ing* suffix, they do successfully recognize verbs in *–ing* contexts (Willits, Seidenberg, & Saffran, 2014). In this line of research, I used corpus analyses to understand the structure in the input, used models to make predictions about behavior, and tested these predictions in experimental studies. This led to novel insights about the units infants use during language processing.

Future Directions

Over the next five years, I have a number of plans to extend my research program in ways that capitalize on its strengths. I will continue to make use of my three-way combination of (1) statistical analyses of large, naturalistic datasets, (2) use of those statistics to build theory-driven computational

models of learning and knowledge representation, and (3) testing of those theories in behavioral experiments. I will continue to concentrate in my focus areas, like the organization of children's semantic memory, the interaction of meaning and structure in language acquisition and use, and the identification of the units over which children are doing their computations and performing statistical learning. There are a number of specific directions in which I would like to advance my research program.

First, in accordance with my focus on the analysis of large, naturalistic datasets, one priority is the gathering of larger, audio and visual corpora of child-directed speech. I am especially motivated to increase the diversity of the data, in terms of issues like as socio-economic status, dialect, and bilingualism. My initial work (as well as related work by others) in this domain has shown how powerful and potentially paradigm-shattering these large datasets will be, and I plan to continue to be a leader in using these techniques to better understand language acquisition and related phenomena.

Second, in accordance with my focus on building theory-driven computational models, we are developing a number of formal computational models of the development of semantic memory, the acquisition of syntactic structure, and how these kinds of knowledge bootstrap each other's acquisition. One specific future direction for this work is the construction of models of human neural and brain development, and the impact of these processes on language acquisition. Some of this work is being done in the form of new and exciting collaborations. One such collaboration is with Dr. Jochen Triesch at the Frankfurt Institute for Advanced Study of Neuroscience, involving computational neuroscientific models of knowledge acquisition. A second collaboration is with Dr. Harald Baayen and Dr. Michael Ramscar at Tubingen University, studying adaptive learning models of language acquisition. A third collaboration is with Dr. David Landy of Indiana University, modeling how language impacts children's acquisition of number concepts.

Third, in accordance with my focus on testing these models in behavioral experiments, we are collecting more and better experimental data from infants, toddlers, children, and adults, and their learning and language development. One concrete plan is to map out the developmental trajectory of the structure of children's semantic memory, using experimental paradigms like EEG as well as the semantic priming paradigm I previously developed for infants (Willits et al., 2013). An important component of this work will be the continued development of novel behavioral paradigms, including "games" for children we have been developing for smartphones and tablet computers. These methods will be an important component of grants focusing on testing corpus and model-based hypotheses about how children learn and represent knowledge about word meaning, semantic structure, and other aspects of language acquisition.

Finally, an important component of my work is the continuation of recent work that applies models of naturalistic data to translational questions involving clinical and special populations. One such project is a study of the vocabulary and semantic development in profoundly deaf children with cochlear implants. We have been creating a model of the development of semantic knowledge in deaf children with cochlear implants by taking our model of typical children's semantic development and making it more reflective of the experiences of these children (Willits, Jones, Pisoni, & Kronenburger, in preparation). We then use the model to predict outcome measures such as vocabulary development and behavior in various experiments testing the development of semantic knowledge. Another such project done in collaboration with Dr. Paul Lysaker of Indiana University, involves studying individuals with schizophrenia. In this work, we are using recordings from naturalistic settings as well as clinical sessions to try to predict symptoms, disease progression, and responsiveness to various treatments by measuring various factors of the individuals language use.