Chapter 1: A Theory of the Language Learning Mechanism

1.1 The Mechanism of Language Learning

Language learning is a curious enterprise, effortless for children while often effortful for adults. This intriguing dichotomy has been the subject of intense research in linguistics and psychology, and this dissertation focuses on how children could accomplish the difficult task of language learning with such unconscious ease.

Understanding the mechanism of language learning is vital once we consider the complexity of the system to be learned. Like many other systems, the linguistic system is comprised of many different pieces. In addition, again like many other systems, the linguistic system often has a non-transparent relationship to the observable data points generated by it, which is what a learner has access to. Both of these conspire to make language learning a non-trivial undertaking.

One way to address this problem is to constrain the systems the learner could acquire by defining a finite set of parameters the learner must set in order to "learn" the language(s) of the surrounding environment (as in Chomsky (1981), among many others). This serves to ease the learner's burden since only systems with particular features will be considered. However, this does not *solve* the problem of language learning. Suppose, for example, that the potential systems a learner could acquire are described by *n* binary parameters. This still leaves 2^n possible systems for the learner to choose from, which is a large number indeed (as noted by Clark (1994), among many others) even for *n* as low as 10 or 20. The problem remains of how the learner chooses from among that set of potential systems, given the observable data which is often highly ambiguous and exception-filled. This is what a theory of the mechanism of language learning endeavors to explain.

Investigation of the language learning mechanism requires knowledge of both the system to be acquired and the time course of acquisition. Theoretical linguistics can provide a description of the object of acquisition, which is the linguistic system that adults use and children must acquire. Experimental research can furnish the milestones of acquisition: by a certain age, children behave as though they know certain pieces of the linguistic system. Given these two boundary conditions - the linguistic representations and the trajectory of language learning - we can then explore the means by which learners could acquire pieces of the system in the time frame that they do.

1.2 Language Development: Constraints on the Hypothesis Space

From the biological perspective, the development of language is an interaction between internal and external factors (Yang, 2002; Baker, 2001; Lightfoot, 1982; among many others). One interpretation of internal factors would be as constraints on the hypotheses under consideration by the learner. The most prominent instantiation of such constraints are linguistic parameters (Chomsky, 1981), though there are other ways the learner's hypotheses might be constrained. It is, however, crucial that the learner's hypothesis space be defined by the time the learner is attempting to decide *which* hypothesis is correct for the exposure language.

The hypothesis space may be defined in terms of parameters, with one parameter value per hypothesis (as in Yang (2002)). But the hypothesis space does not *have* to be defined this way; for instance, the learner might instead have a hypothesis space defined over the amount of structure posited for the language: linear vs. hierarchical (see, for example, Perfors, Tenenbaum, & Regier (2006)). The key point is that the learner's hypothesis space is defined, however that may be instantiated. External linguistic experience will then shift the learner's beliefs in the various hypotheses under consideration.

1.3 Formalizing the Language Acquisition Mechanism

The language acquisition process has been described formally by Yang (2002), using three components: a language learning algorithm L, a set S of potential states the learner can be in, and experience from the linguistic environment E. The learning algorithm L takes the initial state S_0 of the learner, which includes a defined hypothesis space of the linguistic structures under consideration, and updates it with external linguistic experience E until the learner reaches the target state S_T.

(1) $L(S_0, E) \rightarrow S_T$

When the learner is in S_T, the learner has acquired the adult system of linguistic knowledge. The learning algorithm L encapsulates the mechanism of language learning, as it is the procedure by which the learner converges on the appropriate linguistic hypothesis (formalized as the learner being in state S_T) by the appropriate time. However, there are sub-components of L that can be made explicit. In addition to a procedure to update the learner's beliefs about the correct hypothesis, L should also include a procedure that decides which data to learn from (the data *intake* (Fodor, 1998b)).

The entire learning framework thus consist of three parts: (1) a definition of the hypothesis space, (2) a definition of the data intake, and (3) a definition of the algorithm that searches the available hypotheses and, based on the intake, converges on the correct one(s). We can easily map these framework components to the formal definition components described previously. The definition of the hypothesis space is part of the definition of the learner's initial state S_0. The data intake and update procedure are captured in the learning procedure L.

1.4 Domain Specificity and Domain Generality

Defining the learning theory in this somewhat abstract manner allows us to apply it to a range of learning problems. In addition, we can combine discrete linguistic representations (the defined hypothesis space) with probabilistic methods (the update procedure). This is a quite a useful outcome, as linguistic representations are often associated with domain-specific knowledge while probabilistic methods are often associated with domain-general knowledge and the debate has long raged over whether language learning is domain-general or domain-specific.

Dividing the learning theory into three components allows us to examine them separately, and importantly allows for a learning theory that can be both domain-specific and domain-general. Thus, this framework allows for a synthesis of the two approaches, retaining the positive benefits of each. Learners may be constrained in the representations that comprise the hypothesis space, the data they deem relevant for learning, or the procedures they use to update their beliefs about the available hypotheses.

1.5 Investigating the Components of the Learning Framework

Each of the components of the learning framework can be investigated separately. The question of exactly how the hypothesis space is defined, for instance, has been the source of vast amounts of spilled ink and hard feelings. Scores of theoretical and experimental work (Chomsky, 1981; Hamburger & Crain, 1984; Thornton & Crain, 1999; Lidz, Waxman, & Freedman, 2003; among many others) have been dedicated to identifying what hypotheses children entertain at given points in time, how they are constrained in what hypotheses they initially consider, and how they are constrained in what hypotheses they might later posit. Recently, experimental work has also been devoted to investigating the updating procedure, instantiated as a domain-general statistical updating procedure akin to Bayesian updating. Based on the psychological evidence for such a probabilistic updating procedure in adults (Thompson & Newport, 2007; Bonatti et al., 2005; Newport & Aslin, 2004; Tenenbaum & Griffiths, 2001; Cosmides & Tooby, 1996; Staddon, 1988), recent experimental work has tackled the existence of a similar probabilistic procedure in young language learners (Gerken, 2006; Gerken, 2004; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; among many others).[1]

We can look also to the data intake filtering component. Intuition about how learners might behave leads us in two opposite directions. On the one hand, using all available data could uncover a full range of patterns and variation. This is especially true from the viewpoint of statistical modeling. Probabilistic models are often inhibited by sparse data (in fact, many smoothing techniques exist precisely for this reason (Jurafsky & Martin, 2000; Manning & Schütze, 1999)), so any truncation of the data set available for language acquisition seems ill-advised. On the other hand, the observable data is noisy. Perhaps data that are more transparently related to the underlying linguistic system (more "informative" or more easily "accessible" data)

[1] Note that the learner's ability to track probabilities does not negate the need for constraints on the hypothesis space. Some experimental work on young language learners in fact supports constraints on the hypotheses the learner considers. Specifically, Gerken (2004) show that infants can induce an abstract generalization from data that does not exhaustively signal this generalization. In order to do this, the hypothesis space containing that abstract generalization must already be defined. Learners must posit (and analyze data for) that specific generalization as opposed to however many other generalizations are compatible with the observed data. Gerken (2006) demonstrates that infants have a preference for making a more restrictive generalization when two are available. In order to do this, the hypothesis space has to already be defined – one hypothesis for the more restrictive generalization and one hypothesis for the less restrictive one. So, probabilistic learning is a procedure that is used once the hypothesis space is constrained to those two hypotheses. Probabilistic learning is *not* an alternative to defining the hypothesis space.

are easier for the learner to extract the correct systematicity from. Thus, even though such data would be significantly sparser, they would lead to the correct generalizations about the underlying system that produced the observable data.

1.6 Computational Investigations of Data Intake Filtering

In the current work, I examine several language learning case studies that suggest children must filter their data intake down to a more informative and accessible (if sparser) subset of the available data. Key to this work is the exploration via computational modeling of both synchronic and diachronic data, since the most direct experimental technique of testing filtered data in a naturalistic environment is logistically (and ethically) difficult to implement. We would have great trouble restricting the intake of a young child (let alone a whole group of young children) for an extended period of time and seeing the effect of this restriction on the acquisition of the target language. For simulated learners, however, this restriction is quite simple. It is perfectly feasible to restrict the data intake of a simulated learner in any way we choose and then observe the effect on the model's learning.

One question that might reasonably arise is how much use a simulated learner actually is. Why do we believe that a model of a learner is at all realistic? As Goldsmith & O'Brien (2006) note:

> "When the model displays unplanned (i.e. surprising) behavior that matches that of a human in the course of learning from the data, we take some satisfaction in interpreting this as a bit of evidence that the learning models sheds light on human learning."

In short, if the simulated learner accords with human behavior in some non-trivial way that is not purposefully built into the model, we conclude that the assumptions the learning model has made accord with the human learning algorithm. And indeed, there has been a recent surge of computational modeling work examining the effect of data filtering on language acquisition (Sakas & Fodor, 2001; Sakas & Nishimoto, 2002; Yang, 2002; among others).

This dissertation continues the nascent computational modeling tradition by investigating data intake filtering in three separate case studies covering different learning problems in various domains of linguistics: the syntax-semantics interface, syntax, and metrical phonology. In each case, the hypothesis space is defined using domain-specific hypotheses and the update procedure is an adapted form of the domain-general procedure of Bayesian updating. With these two components set, we can then investigate the effects of the remaining component: data intake filtering.

1.7 Organization of Dissertation

The dissertation proceeds as follows:

Chapter 2 describes the adaptation of Bayesian updating to a linguistic framework, specifically a hypothesis space with two pre-specified hypotheses. This chapter is meant as a primer to the mathematical underpinnings of the update procedure that will be assumed in the subsequent chapters.

Chapter 3 examines the case of learning anaphoric *one* in English, a language learning problem that spans the domains of structure and reference in the world. Experimental evidence has suggested that children have acquired this knowledge by 18 months (Lidz, Waxman, & Freedman, 2003) and I explore how a child could accomplish this feat, given realistic estimates of the data available to children. Based on the learning models results, I argue that data intake filtering is a necessary part of successful acquisition of English anaphoric *one*.

Chapter 4 explores a scenario where the adult target state is a probability distribution between two hypotheses. This was the case for Old English word order between 1000 and 1200 A.D. Under the assumption that the Old English shift from Object Verb to Verb Object order is due to misconvergences on the correct target probabilities during learning (Lightfoot, 1991), I implement a model of Old English language change for a population of individuals that use a particular learning algorithm. Correct population-level behavior only results when individuals filter their data intake during learning in specific ways. This case study serves as a second argument for the necessity for data intake filtering, in addition to the feasibility of data intake filtering in a realistic system.

Chapter 5 investigates how a child could learn English metrical phonology. This is a difficult task as the system is complex, involving 9 interacting parameters (Dresher, 1999), and the observable data from the target language is extremely noisy. For this scenario, we can examine the feasibility of data intake filtering in a truly hard learning environment. I examine two methods of implementing a specific data intake filter, and demonstrate that both methods can lead to successful acquisition. The ability to solve the language acquisition problem for the complex, noisy system of English metrical phonology is again support for the feasibility and sufficiency of data intake filtering.

Chapter 6 summarizes the main points from the case studies examined in the dissertation and highlights the contributions from this dissertation to linguistics, learnability, and computational modeling.