Testing the Universal Grammar Hypothesis

1. Introduction

Perhaps the single most controversial claim in linguistic theory is that children learning their native language face an induction problem, or in other words, that the available input underspecifies the adult state. This induction problem is known by many names: the "Poverty of the Stimulus" (e.g., Chomsky 1980a, Chomsky 1980b, Lightfoot 1989, Crain 1991), the "Logical Problem of Language Acquisition" (e.g., Baker 1981, Hornstein & Lightfoot 1981), and "Plato's Problem" (e.g. Chomsky 1988, Dresher 2003). Regardless of the name, it all boils down to the same claim: the data generally available to young children are compatible with multiple hypotheses, or perhaps more correctly, data necessary to rule out the incorrect hypotheses are either not available at all, or not available in sufficient quantity (Lightfoot 1982, Legate & Yang 2002, among many others).

The Universal Grammar (UG) hypothesis was introduced as a solution to this problem (Chomsky 1957/1975, 1965). The logic of the UG hypothesis is straightforward: if the necessary evidence for choosing the correct linguistic hypothesis is unavailable in the input, then children must bring some internal bias to the language learning problem (Chomsky 1981, Hornstein & Lightfoot 1981, Legate & Yang 2002, among many others). While the necessity of some kind of bias is generally granted by even the most ardent critics of the UG hypothesis (e,g, Pullum & Scholz 2002, Regier & Gahl 2004), the nature of the necessary biases is the subject of considerable debate. First, there is the question of what cognitive objects the bias operates over. A bias might operate over the representations the child considers as hypotheses (e.g., parameters of linguistic variation (Chomsky 1981)), the data the child learns from (e.g., only unambiguous data (Fodor 1998)), or the learning algorithm the child uses to alter belief in competing hypotheses (e.g. trigger-based learning (Gibson & Wexler 1994, Niyogi & Berwick 1996)). Second there is the question of whether the necessary bias is specific to language learning (i.e. domain-specific) or applies generally to any kind of cognitive learning (domain-general). UG is usually proposed as a collection of domain-specific biases ranging over both the representations that children consider and the data that children learn from, but this is far from logically necessary (e.g., Chomsky 1971, 1981, Kimball 1973, Baker 1978, Gordon 1986, Lightfoot 1991)

Recent debates about the UG hypothesis have tended to focus on two broad questions. The first concerns the existence of the induction problem (e.g., Sampson 1989, 1999, Pullum & Scholz 2002, MacWhinney 2004, Tomasello 2004), which is, of course, the motivation for the UG hypothesis. Until recently, the claim that children's input lacks sufficient evidence for successful language learning has been based on the intuitions of linguists rather than on large-scale empirical analyses of child-directed speech. However, without quantifiable evidence for induction problems, there is no need for the UG hypothesis at all. In fact, Pullum and Scholz (2002) have claimed just that: using data from the Wall Street Journal corpus (Linguistic Data Consortium 1993) and the CHILDES database (MacWhinney 2000), they argue that there is no evidence for an induction problem for several well-known linguistic phenomena in English such as anaphoric *one* (Baker 1978) and yes-no questions involving complex subjects (Chomsky 1971). Even granting the existence of an induction problem for a given linguistic phenomenon, the second broad question follows directly: what is the nature of the prior knowledge necessary to solve that problem? More specifically, is the knowledge innate or derived from prior learning? Is the knowledge domain-specific or domain-general? One could imagine any or all of the possible combinations being applicable to various aspects of the linguistic system: some knowledge may be innate and domain-general, some innate and domain-specific, some derived from domain-general knowledge acquired previously, and some derived from domain-specific knowledge acquired previously. With the proliferation of possible types of prior knowledge, it is not clear that a single type will be sufficient to solve all of the induction problems in language learning. In fact, Tomasello (2004) takes this one step further: he argues that the proliferation of specific suggestions for that prior knowledge in the theoretical literature has rendered the UG hypothesis

untestable through standard scientific falsification. He contends that it will not be possible to evaluate the UG hypothesis until it is broken down into specific hypotheses about biases with respect to specific linguistic phenomena.

The project we propose here aims to address both of these questions directly, and in the process lay out a concrete methodology for testing the UG hypothesis that is in similar in spirit to what both the critics of the UG hypothesis (e.g., Pullum and Scholz 2002 and Tomasello 2004) and the supporters of the UG hypothesis (e.g., Chomsky 1957/75, Crain and Pietroski 2002) propose. Utilizing techniques recently made possible through advances in technology, and combining aspects of theoretical, experimental, and computational linguistics, it is now feasible to perform several quantitative tasks relevant to evaluating the UG hypothesis with respect to the issues discussed above. We can search reasonably large corpora of both adult and child-directed speech for relevant linguistic structures; we can precisely measure the adult knowledge state children eventually attain using psycholinguistic techniques from experimental syntax; and we can implement sophisticated probabilistic learning models (specifically Bayesian models) capable of operating over the structured representations postulated by linguistic theory. With these techniques in hand, we plan to investigate the existence of the induction problem by examining both the realistic data used as input by children (available through resources such as CHILDES (MacWhinney 2000)) and the knowledge state achieved by adults for complex linguistic phenomenon such as syntactic islands (e.g., the experiments in Sprouse 2007). We will then implement Bayesian learning models to test whether unbiased learners can reach the adult knowledge state given the data available. If unbiased learners cannot do this, then we can conclude that the induction problem does indeed exist for that phenomenon and that children require learning biases to succeed. We can then identify what kind of biases lead to acquisition success by incorporating different types of learning biases into the models (as is done, for example, for learning anaphoric *one* in Pearl & Lidz (submitted)). The biases implemented may be domain-general in nature (e.g., Regier & Gahl 2004, Perfors, Tenenbaum, & Regier 2006, Pearl & Lidz submitted) or domain-specific (Sakas & Fodor 2001, Pearl & Weinberg 2007, Pearl 2008, submitted, Pearl & Lidz submitted). Crucially, because the Bayesian modeling framework allows us to accommodate biases of many kinds, from choosing the smallest hypothesis consistent with the data (Tenenbaum & Griffiths, 2001) to restricting the input to certain clauses (Lightfoot 1991, Pearl & Weinberg 2007) to constraining the representations under consideration via parameters (Chomsky 1981), we will be able to both reduce the UG hypothesis to smaller specific hypotheses and evaluate the necessity of those hypotheses for successful learning (for instance, as advocated for by Tomasello (2004)).

2. Accurate measures of the primary data

The first step of our investigation is to assess the input that is actually available to children for various linguistic phenomena. Since the debate regarding the induction problem and the necessity of UG hinges on the state of children's input, occurrence facts about child input should not be based on the intuitions of linguists (an idea advocated extensively in Pullum & Scholz (2002), for instance). This is particularly true now that corpora of child-directed speech are freely available, such as CHILDES (MacWhinney 2000). Notably, however, the corpora available are rarely marked with all the information of interest to a linguist focused on complex syntactic and semantic phenomena, which are primarily the locus of the induction problem debate (Crain & Pietroski 2002, Legate & Yang 2002, Pullum & Scholz 2002, Lidz, Waxman, & Freedman 2003, Reali & Christiansen 2004, Regier & Gahl 2004, Kam et al. 2005, Perfors, Regier, & Tenenbaum 2006, Foraker et al. 2007, Pearl & Lidz submitted, among many others). While some corpora may contain morphological information or part-of-speech identification, most are simply transcripts of child-directed speech. We propose to annotate several available child corpora in the CHILDES database syntactically (using, for example, the features in Government and Binding Theory (Chomsky, 1981)) via a two-step process. The output of this process will be fully formed

hierarchical structures, so that formal analyses from theoretical linguistics can be easily adopted as biases in the models we later build (see sections 4 and 5 for details). First, we will use a freely available dependency tree parser (such as the Charniak parser[1]) to generate a first-pass syntactic analysis. Then, we will evaluate the resulting syntactic trees by hand (with the help of undergraduate research assistants), correcting when necessary, to ensure the accuracy of the structures generated. We intend to make the final parsed corpora available through CHILDES for other language researchers to use.

In addition, we propose to investigate adult corpora of conversational speech (such as those available through TalkBank (http://www.talkbank.org) in order to compare the differences between adult and child-directed speech for various linguistic phenomena. Often, child-directed speech corpora are relatively sparse compared to available adult speech corpora, especially if syntactic annotation is desired, which has led much of the corpus-based linguistic research to rely on adult-directed speech (e.g., Pullum & Scholz (2002)). Yet, it is a common (and quite reasonable) argument that child-directed speech may differ quite significantly from adult speech (see, for example, discussion in Legate & Yang (2002)). Given that recent probabilistic learning models are sensitive to the relative frequencies of various data (e.g., Foraker et al. 2007), it seems only prudent to ask, for a given linguistic phenomenon, if the data frequencies do differ. It may turn out for some linguistic phenomena that the relative frequencies do not vary much between the speech directed at, say, three-year-olds and the speech directed at adults. This would then suggest that adult speech corpora may indeed be a reasonable estimate of children's input for some phenomena, particularly complex syntactic and semantic interpretation phenomena that are acquired later in development (e.g., negative polarity items like 'any', the interpretation of connectives such as 'or', and binding theory phenomena, as discussed in Crain & Pietroski (2002)). Given the abundance of adult-directed conversational speech, such a scenario would provide a far richer source of data from which children's input could be estimated. However, should child-directed and adult-directed speech frequencies differ, it will be crucial to this project to determine not only if, but also in what way they differ, so as to correctly evaluate both our own models and those potentially offered by others.

Like the child-directed speech, much conversational adult-directed speech is not annotated with syntactic information. The process we propose to use to generate annotated adult-directed speech corpora is identical to the process for generating the annotated child-directed speech, involving a first-pass annotation by a freely available parser and subsequent human evaluation of the generated annotation. We intend to make the annotated corpora available to the research community either through TalkBank (http://www.talkbank.org) or the Linguistic Data Consortium (http://www.ldc.upenn.edu/), a common repository for electronic corpora.

3. Accurate measures of the adult state

The second step of our investigation is to assess the adult knowledge state children eventually attain. It almost goes without saying that acceptability judgments form the primary measure of the adult grammar in the field of theoretical syntax; therefore, acceptability judgments are the logical choice for a quantifiable measure of the adult state. There are at least three reasons for the predominance of acceptability judgments in the study of adult grammars. First, acceptability judgments can be provided with little effort from the subject (Schutze 1996, Cowart 1997). Second, these judgments are highly reliable across speakers of the same language (Cowart 1997, Keller 2000, Sprouse 2007). Third, these judgments are a robust proxy for grammaticality (Chomsky 1965, Schutze 1996, Cowart 1997, and many others). Paradoxically, the very properties that have made acceptability judgments such a valuable data source for theoretical syntacticians have also served to undermine general confidence in that data. First, because

[1] Available through Brown University (ftp://ftp.cs.brown.edu/pub/nlparser/).

judgments are available to any native speaker, linguists have tended to use *their own* judgments rather than those of naïve consultants (Christiansen and Edelman 2003). Second, because judgments are generally reliable across speakers, linguists have tended to use single data points rather than samples (Bresnan 2007, Cowart 1997). Third, because judgment tasks are often designed as a choice between grammatical and ungrammatical, until recently relatively little research has been done on the gradience inherent to acceptability judgments, and the factors that might be causing or influencing that gradience (Keller 2000, Sorace and Keller 2005).

In response to these concerns, several linguists have developed a set of formal methodologies, which have collectively come to be known as *experimental syntax*, for collecting acceptability judgments. While the details vary from experiment to experiment, experimental syntax methodologies all have at least four components in common (Featherston 2007, Sprouse 2007). First, judgments are collected from a sample of naïve consultants, usually at least 10 and ideally more than 20, to insure that judgments generalize to the broader population. Second, consultants are presented with a variety of sentences for any given structure under investigation, to insure that the judgments generalize across lexical items. Third, consultants are presented with a formal task, such as a Likert Scale task or the Magnitude Estimation task (Stevens 1957, Bard et al. 1996), to help insure that relative acceptability data are not lost to categorical responses. Fourth, data are analyzed using standard behavioral statistics. For this project, we will use experimental syntax techniques to measure the relative acceptability of structures in the adult grammar for comparison to the relative frequencies of those structures in the child-directed speech corpora and adult conversational speech corpora.

Experimental syntax methodologies have advantages over previous informal collection techniques too numerous to mention here (see Schutze 1996, Cowart 1997, Keller 2000, Featherston 2007, and Sprouse 2007 for discussion). However, given the nature of this project - in particular, the comparison between relative frequencies and acceptability judgments - two of these advantages bear mention. First, experimental syntax has introduced rating tasks, such as magnitude estimation (Stevens 1957), that provide a more precise measure of relative acceptability than previous informal collection tasks. Most informal collection tasks involved binary rating scales such as *yes/no* or limited, discrete rating scales such as the 5 or 7 point Likert scales. All of these limited scales can result in a loss of information to categorization (Bard et al. 1996). In contrast, magnitude estimation places no predefined restriction on the response scale: subjects may use the entire positive number line for their responses, thus eliminating the categorization problem. Bard et al. (1996) demonstrated that given such freedom, subjects routinely distinguish more than 7 levels of acceptability. Furthermore, Sprouse (submitted b) has demonstrated that subjects' responses in magnitude estimation tasks are incredibly robust across samples, even with minor variations to the experimental design (such as modifying the modulus sentence). Taken together, these facts suggest that newer rating tasks such as magnitude estimation will provide more detailed data regarding the adult grammar.

Second, experimental syntax has also introduced the principles of factorial experimental design, which has enabled the investigation of contributions from factors that are traditionally outside the domain of syntactic theory, but that may still have an effect on both acceptability judgments and (crucially) relative frequencies. For example, Sprouse (2008, submitted a) both demonstrate that the acceptability of wh-movement dependencies is affected by the distance of the dependency (see also Frazier (1989) and Phillips et al (2005)). Specifically, shorter wh-movement dependencies (1) are significantly more acceptable than longer wh-movement dependencies (2) despite the fact that syntactic theories predict both structures to be categorically grammatical.

(1) Jack hoped that you knew **who** the giant would chase.
(2) Jack knew **who** you hoped that the giant would chase.

Furthermore, Sprouse (2007, submitted a) both demonstrate that the acceptability of wh-movement dependencies is also affected by the complexity of the structures involved. Simpler structures (3) are more acceptable than more complex structures (4); this is again true despite the fact that syntactic theories predict both structures to be categorically grammatical.

(3) Who thought that Jack climbed the beanstalk?
(4) Who wondered whether Jack climbed the beanstalk?

It would not be surprising to find that factors like dependency distance and complexity also affect the relative frequency of these structures. It is likely that these acceptability effects reflect the processing load required for these structures (Kluender and Kutas 1993a,b, Phillips et al 2005), and that there is a correlation between processing load and frequency of production. In this vein, section 5 reports a pilot study of these factors in an investigation of syntactic island effects (Ross 1967), the results of which suggest just such a correlation between extra-grammatical acceptability effects and relative frequency.

4. Probabilistic Models of Language Learning

Once we have an accurate assessment of both the input available to children and the target state for children (in the form of the adult knowledge state), we can then test ways in which children could use the available data to reach that target state. This is where computational modeling comes into play. Computational models enable us to test the learnability of a particular piece of linguistic knowledge very precisely. To put it simply, modeling can be used for any linguistic problem where there is a theoretical claim about learnability, a defined input set, and a defined target state. Here, both the input set and target state are defined. The learnability claim concerns what kind of biases (if any) are necessary to reach that target state, given the input.

If the adult state cannot be achieved without some kind of learning bias – be it over the representations, data intake, or learning algorithm- then the language learning induction problem does indeed exist. Then, computational modeling allows us to explore what biases make the target state achievable from the available data. So, through modeling, we can accomplish three goals. First, we can test the motivation for UG. Second, we can identify what learning biases are necessary. Third, we can determine whether those biases are domain-specific and not derivable from prior knowledge - and so part of UG.

The strength of computational modeling lies in its ability to create a language acquisition mechanism that we have complete control over. In this way, we garner data about learnability that would not otherwise be available. However, the point of modeling is to increase our knowledge about the way that *human* language acquisition works, not simply provide a computational or mathematical model capable of solving a particular problem. So, we must be careful to ground our model empirically – that is, we must consider if the details of the model are psychologically plausible by looking at the data available on human language acquisition. In particular, we should consider whether the learning algorithm is realistic. Probabilistic reasoning (particularly Bayesian) has been shown to be the optimal strategy for solving problems and making decisions given noisy or incomplete information (Pearl 1988). There is also evidence for the psychological validity of a procedure like Bayesian learning as a method used by adult humans (Staddon 1988, Cosmides & Tooby 1996, Tenenbaum & Griffiths 2001), young children (Xu & Tenenbaum 2007), and infants (Gerken 2006). But it is important to keep in mind that reasoning, probabilistically or otherwise, requires an adequate understanding of the representations that are used in the relevant mental computations. A probabilistic learner takes probabilistically available information to derive a conclusion about a discrete representation from a range of antecedently available options (cf. Shannon 1948). Crucially, for probabilistic learning such as Bayesian learning to be able to function, the hypothesis space must already be specified (cf. Tenenbaum,

Griffiths, & Kemp (2006) for theory-based Bayesian models that emphasize this point). Otherwise, Bayesian learning has nothing over which to operate. For this reason, probabilistic learning – and Bayesian learning in particular – is quite effective for examining the learnability of explicitly defined linguistic knowledge where a range of options is already available to the learner (e.g., syntactic parameters and hierarchical structure (Chomsky 1981)).

Recently, there have been several probabilistic learning models developed for what have been considered classic examples of the induction problem in language learning (e.g., structure-dependent learning as exemplified in complex yes-no question formation: Reali & Christiansen 2004, 2005; Kam et al. 2005, Perfors, Tenenbaum, & Regier 2006; and anaphoric *one* interpretation: Regier & Gahl 2004; Foraker et al. 2007; Pearl & Lidz submitted). While some do not operate over the structured representations that are intrinsic to linguistic theory (e.g., Reali & Christiansen 2004, 2005), those of the Bayesian persuasion do (e.g., Regier & Gahl 2004). This gives the Bayesian models a distinct advantage in terms of speaking to the induction problem linguists traditionally envision. However, many recent Bayesian models are presented in an "ideal learner" framework, where the model does not learn incrementally as children do (see Perfors, Tenenbaum, & Regier (2006), Foraker et al., (2007), among others). Instead, the model performs computations over the entire input set at once, which would be equivalent to children remembering everything they ever heard and analyzing it all at once – a mental feat unlikely for adults, let alone children. However, these idealized models do demonstrate that the necessary information could be available to children in the data - if that information can be accessed - without recourse to any additional prior knowledge.

While incremental forms of these Bayesian models are possible (see Doucet, De Freitas, & Gordon (2001) for several examples), they have often not been implemented because of the way the induction problem is viewed. Rather than it being taken as a statement that *children* cannot reach the adult knowledge state given their input, it is instead taken as a statement that the inference from input to target state simply *cannot be made* no matter what computational resources are available (Perfors, Tenenbaum, & Regier 2006). Thus, while these models do address the induction problem at the logical level and admirably operate over the structured representations linguistic theory trades in, they do not quite connect with the underlying spirit of the induction problem. The heart of the induction problem revolves around what children can do, not ideal learners with infinite memory. To this end, there have also been recent attempts at incremental Bayesian models for induction problems, such as those dealing with learning the syntactic category and semantic reference of anaphoric *one* (Regier & Gahl 2004, Pearl & Lidz submitted) and learning the referents of words in general (Xu & Tenenbaum 2007).

One of the insights from recent modeling studies has been that, for a given linguistic phenomenon, the input children learn from may not simply be restricted to the data directly related. For example, English children must learn the correct formation of English yes/no questions that have a complex subject (e.g., *Is [the girl who can solve the labyrinth]*$_{ComplexSubj}$ t_{is} *going to defeat the king?*). Data directly related would be examples of yes/no questions in the input, especially those containing a complex subject. However, the correct question formation relies on a more general bias – that of structure-dependency. Information on whether the language is structure-dependent or structure-independent can come from many different data types in the input, not simply complex yes/no questions. Perfors, Tenenbaum, & Regier (2006) demonstrate that there is sufficient indirect evidence available for this general bias to lead an ideal learner to the correct generalization for complex yes/no questions. Note that this is a kind of indirect evidence that is crucially different from indirect *negative* evidence (Chomsky 1981, Lasnik 1987, among many others), which is the absence of expected data. While structured Bayesian models are able to use indirect negative evidence as well, another strength lies in utilizing evidence that would classically be considered unrelated to the learning problem at hand.

More succinctly, the input for various induction problems was traditionally thought to include only unambiguous data for the hypotheses under consideration (e.g., only complex yes/no

questions to learn the formation of complex yes/no questions (Legate & Yang 2002, among many others), and only unambiguous data for anaphoric *one* to learn its interpretation (Baker 1978, Hornstein & Lightfoot 1981, Crain 1991, among many others)). And indeed, this bias is one that has proven to be useful and possibly vital for acquisition success in some cases (learning word order: Pearl & Weinberg 2007; learning stress pattern parameters: Pearl 2008, submitted). However, in other cases, it may be that children use data that are not unambiguous; instead, they learn from indirectly related data types (Perfors, Tenenbaum, & Regier, 2006) and ambiguous data (Regier & Gahl 2004, Yu & Smith 2007, Xu & Tenenbaum 2007, Pearl & Lidz submitted).

More broadly, recent modeling work has provided the necessary framework to explore questions of learnability and learning bias very explicitly. Given estimates of child-directed speech as input and a target output state, an incremental Bayesian model can be implemented that tests whether the target state is reachable. Two recent examples of this approach are Pearl & Weinberg (2007) for learning word order and Pearl & Lidz (submitted) for learning anaphoric *one*. These are discussed in turn below.

Pearl & Weinberg (2007) (henceforth P&W) examines the alteration between Object-Verb (OV) and Verb-Object (VO) order in Old English. In this case, the final state for adults in Old English is argued to be probabilistically distributed between the two orders (Bock & Kroch 1989, Kroch & Taylor 1997, Pintzuk, 2002). This is in contrast to a final state where only one competing hypothesis is accessed (i.e. only one order used) by adults, which is the scenario typically considered for acquisition. Precisely because the target state is *not* an endpoint (either all OV or all VO word order), it is more difficult to gauge learning success. How close does the learner have to get to the adult probability distribution in order for learning to be deemed successful?

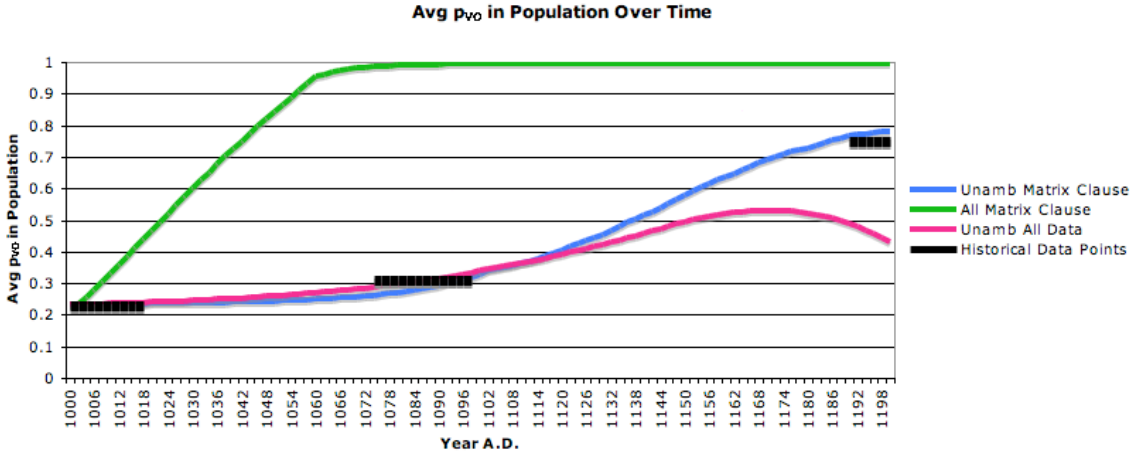
To answer this, P&W make use of the fact that languages change over time. Specifically in the case of Old English, the population shifts from an OV-biased distribution around 1000 A.D. to a VO-biased distribution around 1200 A.D. (PPCME2 Corpus, Kroch & Taylor 2000, YCOE Corpus, Taylor et al. 2003). It has been proposed that certain types of change (in particular, the shift in Old English) result from a misalignment of the child's hypothesis and the adult's analysis of the same data (Lightfoot 1991, 1999). In other words, language change in this case results from *imperfect* learning of a very particular kind. The idea is that language change in this case occurs because learners misconverge on the probability distribution; the learner's probability distribution is very slightly different from the adult's probability distribution. The key point is that the amount of difference between the learner's probability distribution and the adult's probability distribution will influence the rate of language change in a population over time. In order to model change at an attested pace, the acquisition model at the individual level must hypothesize exactly the right amount of difference between the learner's and adult's probability distributions.

So, "successful" learning is defined as learning that leads to exactly the right amount of misconvergence within the individual learner. This amount of misconvergence within the individual then leads to language change over time within the population of individuals. P&W find that the amount of misconvergence depends greatly on how the input is filtered during learning. In this way, they test proposals about data filtering by using models of language change.

The two filters investigated in P&W bias learners away from potentially misleading ambiguous data in the input. Both biases stem from a presumed preference for "simple" data (Lightfoot 1991, Fodor 1998, Dresher 1999, Lightfoot 1999). These biases use a structurally-based notion of simplicity. The first claims that children learn only from unambiguous data (Fodor 1998, Dresher 1999, Lightfoot 1999), and consequently do not learn from data perceived as ambiguous. The second proposal restricts learning to the data points found in "simple" clauses (Lightfoot 1991), where simple clauses are defined as matrix clauses. If there are available data points in embedded clauses, these data are effectively ignored by the learner.

P&W draw on historical data in order to empirically ground their model and calculate the desired rate of change from OV to VO word order in the population. The population model creates a set of successive generations of Old English speakers, each diverging from the initial distribution to a designated extent; this is the model's rate of change and could be compared against the historical rate of change. Each speaker within the population contributed to the data encountered by new population members, who then had to learn the probability distribution between OV and VO word order via incremental Bayesian learning.

Importantly, individual learners could be implemented with or without the input restrictions. The graph below demonstrates the rate of change with different restrictions on learners' input: only unambiguous matrix clause data, only matrix clause data, and only unambiguous data in both matrix and embedded clauses. The average probability of the VO order (p_{VO}) in the population is plotted over time.

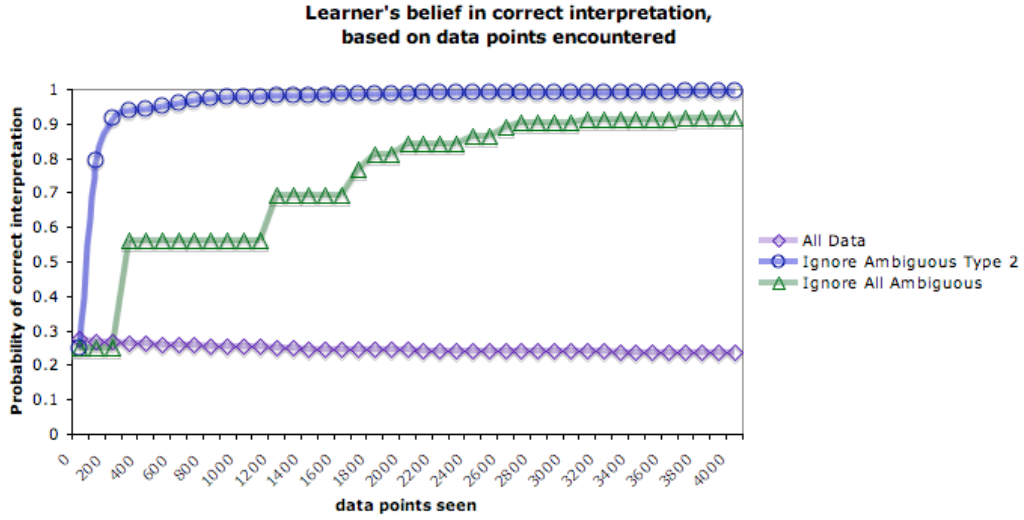


As is apparent, only the population using both input filters (and so learning from only unambiguous matrix clause data) is able to match the data points from the historical records. P&W additionally argue that a population whose individuals learn from all the available data (and who thus have no input restrictions) is also unable to match the historical trajectory. If individuals do not have these biases on the data intake in place, the population as a whole will either shift too quickly or not quickly enough. So, these simulations demonstrate that both input restrictions together allow individual learners in the population to misconverge exactly the right amount at each point in time, which then leads to population-level change at the correct rate.

In sum, P&W find that limiting the data intake is crucial for a successful model of Old English language change. Without certain data intake restrictions, the simulated learners are unable to misconverge the precise amount necessary for the modeled population's rate of change to match the historically attested population's rate of change. Through this computational modeling work, P&W supports the existence of the two proposed filters on data intake during the normal course of syntactic learning.

Pearl & Lidz (submitted) (henceforth P&L) investigates learning the interpretation of the anaphoric element *one* in English. At the syntactic level, the child must learn what the linguistic antecedent of *one* is; at the semantic level, the child must determine what object in the world an NP containing *one* refers to. Both of these levels contribute to the information a Bayesian learner would use when converging on the correct representation of *one*, as a linguistic antecedent (syntax) can be translated into a reference to an object in the world (semantics), and vice versa. In terms of learning, the syntactic antecedent of *one* has semantic consequences on what referents are picked out; the semantic referent has syntactic implications of what the linguistic antecedent is. A probabilistic learner can thus use multiple sources of information to inform its hypotheses about the interpretation of anaphoric *one*.

Like P&W, P&L also uses empirically-grounded computational modeling to examine the effects of learning when the input is either unrestricted or restricted in particular ways. For P&L, the relevant empirical data comes from estimates of child-directed speech to children before the age of 18 months (based on experimental work by Lidz, Waxman, & Freedman (2003)). The probabilistic model used in P&L is a generative Bayesian model that can easily capture learning from various types of data, both unambiguous and ambiguous (see P&L for discussion of the different data types). Importantly, there are two classes of ambiguous data. As the following graph demonstrates, it is crucial for the learner to ignore the second type of ambiguous data in order to converge on the correct interpretation of anaphoric *one* with high probability. The model sees as many data points as children are estimated to encounter before they converge on the correct interpretation (see P&L for discussion of when this happens). If the learner has no input restrictions, then the probability of the correct interpretation is quite low. If instead the learner ignores all ambiguous data, the learner does have a high probability of the correct interpretation after learning – but not as high as when only the second type of ambiguous data are ignored. So, the crucial bias is to ignore only this second type of ambiguous data, rather than all ambiguous data. More importantly, if the learner ignores all ambiguous data, the learner loses a valuable source of indirectly informative data. As noted previously, the restriction of the input set to unambiguous data has led to classic formulations of the induction problem for anaphoric *one*, as the unambiguous data are sparse. So, this modeling work demonstrates the value of selectively learning from more than just the unambiguous data, but crucially not *all* the potentially informative data.



P&L also discusses how a child might come to have just the right bias on input restriction, i.e. to ignore the second type of ambiguous data. They argue that the necessary domain-specific filter (ignore this kind of ambiguous data) can be generated by using a domain-general strategy (learn in cases of uncertainty) provided there is a domain-specific bias on how to view the learning problem (see P&L for detailed discussion).

Overall, P&L demonstrates that a Bayesian learner (and presumably, any probabilistic model) lacking domain-specific filters on data intake will fail to converge on the correct interpretation of anaphoric *one*, given realistic data. So, a necessary bias for learning anaphoric *one* was identified via computational modeling's ability to precisely manipulate the language learning mechanism.

There are several common components in both P&W and P&L. In each case, a realistic input data set was estimated and an output target state identified that represented the adult knowledge state. Unbiased incremental Bayesian models then attempted to reach the target state, and failed. Then, different learning biases were added to the model; in both cases, these were restrictions on

the data intake of the learner (P&W: learning only from unambiguous data (Fodor 1998) in matrix clauses (Lightfoot 1991); P&L ignoring one class of ambiguous data (Regier & Gahl 2004)). In each case, acquisition success was attainable with these learning biases, but crucially not without them. This highlighted the necessity of these learning biases for learning these two linguistic phenomena. The nature of these necessary biases could then be considered. In some cases, they were thought to be domain-general (e.g., learning from unambiguous data (see P&L)); in some, domain-specific (e.g., learning from matrix clause data (see Pearl 2007)); and in some, an intricate interplay of both (e.g., ignoring a certain class of ambiguous data (see P&L)).

In summary, probabilistic learning models provide a viable avenue for examining the learnability of linguistic phenomena. Recent advances in Bayesian modeling underscore how these probabilistic models can be used to test claims of both the existence of induction problems and the nature of the solution to the induction problems that do exist. Specifically, within a Bayesian model, we can examine the learnability of structured representations of linguistic knowledge from realistic data with and without learning biases. In this way, we can pinpoint what induction problems exist and the learning biases necessary to solve them. Once we know the necessary biases for acquisition success, we can then consider what biases are domain-specific and underivable from prior knowledge – and therefore part of UG.

5. Empirical Domain: Constraints on wh-dependencies

As discussed in section 1, one major impediment to productive discussions of the UG hypothesis between nativists and non-nativists is the lack of consensus in identifying induction problems. While the methodologies laid out in sections 1-3 go a long way toward quantifying induction problems, Crain and Pietroski (2002) point out one additional road block: the phenomena that nativists present as strong evidence for the UG hypothesis are rarely the phenomena that non-nativists use to argue against the UG hypothesis (e.g., Pullum and Scholz 2002). To that end, we intend to begin our search for induction problems with a set of phenomena that nativists would unequivocally analyze as requiring UG to successfully acquire: island constraints on wh-dependencies (Ross 1967). Island constraints fit all three of the criteria put forth by Crain and Pietroski for successful UG arguments. First, a complete analysis of island constraints requires access to cross-linguistic data that is unlikely to be available to children (let alone non-linguists). Second, young children nevertheless demonstrate knowledge of these constraints at a very young age (e.g., the wh-island constraint in de Villiers, Roper, and Vainikka 1990). Third, in the few instances that children do create wh-dependencies that are not present in the adult grammar, they only attempt dependencies that are available in an attested human language (e.g., medial wh-words in Thornton (1990)). Furthermore, while these constraints are often grouped into a single family called 'islands', there are in fact several species of islands: subject islands, adjunct islands, wh-islands, noun-complement islands, relative clause islands, sentential subject islands, if/whether islands, negative islands, factive islands, and perhaps others. With so many species of islands, we can investigate a variety of potential induction problems while maintaining the cohesion of a single investigative domain.

As a concrete example, let's consider subject islands. To control for the effects of dependency length and construction complexity on acceptability (Kluender and Kutas 1993a, b, Phillips et al. 2005, Sprouse 2007, 2008), Sprouse (submitted a) argues that islands should be treated as a two-way interaction of these two factors. In the case of subject islands, this results in following four conditions, which cross the length of dependency with the complexity of the NP:

(5) Simple NPs

(a) subject: What do you think __ interrupted the TV show?

(b) object: What do you think the speech interrupted __ ?

(6) Complex NPs

(a) subject: What do you think [the speech about __] interrupted the TV show?

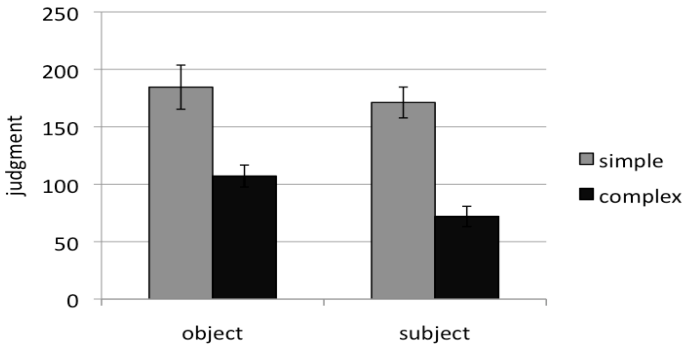
(b) object: What do you think the speech interrupted [the TV show about __]?

With this definition in hand, we can then look at both the relative frequency of these four conditions, and the relative acceptability for adults of these four conditions. As a pilot study, we parsed three adult conversational speech corpora into GB-style syntactic trees using the Charniak parser: the CALLHOME English corpus, the Switchboard Cell corpus, and part of the Fisher-Training-1 corpus (all available through the Linguistic Data Consortium). We then did an automated search for the four structures above, and verified the resulting examples by hand. The results are below in Table 1. Notably, while simple NP extraction examples appear, complex NP extraction examples seem to never be used by adults in these corpora.

Table 1. Results of Pilot Study on Adult Conversational Speech

	Total Utterances	Simple NP subject	Simple NP object	Complex NP subject	Complex NP object
CALLHOME English	2427	73	61	0	0
Switchboard Cell	3630	62	148	0	0
Fisher-Training-1 (part)	22750	600	273	0	0
Total	28807	735	882	0	0

We note that these counts may change for corpora of child-directed speech, but they do provide a rough estimate of the relative frequencies of these four constructions. The question then is whether these frequencies correlate with the acceptability ratings given to these constructions by adult native speakers. To test this, we gave 22 native speakers of English an acceptability judgment task using the magnitude estimation task (Bard et al. 1996) with a reference sentence that was assigned a value of 100. The survey itself tested several island types; however for our present purposes, we will limit the discussion to subject islands. The results are summarized in the graph below. As the graph indicates, there is no difference in acceptability between extraction from a simple object and extraction from a simple subject ($t(21)=.885$, $p=.387$). This mirrors the nearly identical frequencies we obtained for the two simple NP conditions. The graph also indicates a large difference between the simple NP conditions as a group and the complex NP conditions as a group (confirmed by a main effect in a two-way ANOVA $F(1,20)=31.9$, $p<.001$). Again, this mirrors the difference in frequencies obtained from the corpora (1617 versus 0). However, we also find a significant difference between extraction from complex objects and extraction from complex subjects ($t(21)=3.57$, $p=.002$), and an interaction in a two-way ANOVA $F(1,20)=6.41$, $p=.02$). While this is unsurprising from the point of view of syntactic theory – extraction from a complex subject is precisely the definition of a subject island –it is surprising given that we find no difference between the frequencies of these two conditions in the corpora: both had zero



occurrences. This leads to the obvious question: if these constructions have identical occurrence frequencies (in this case, around zero), why is it that speakers report one to be more acceptable than the other? If child-directed speech frequencies also follow this pattern, then this seems a good candidate for an induction problem for children. More specifically, children must require some kind of bias in order to attain the adult knowledge state where complex object extraction is acceptable while complex subject extraction is not.

Nonetheless, it should be noted that this conclusion rests solely on *direct syntactic evidence* for subject islands. As Perfors, Tenenbaum, and Regier (2006) observed, even when there isn't enough direct syntactic evidence for a given constraint, there may be sufficient related evidence to arrive at the correct distributional regularities without relying on additional learning constraints. The question then is what other types of evidence may be used by children to correctly learn subject islands. We believe the analyses in the theoretical literature already provide a road map for answering this question in at least three ways.

First, wh-island constraints may be an instance of more general displacement dependencies, which would underlie structures such as relative clauses, topicalization, and even adjective-though constructions (e.g., Smart though Bill is, he can't tie his own shoes.). A plausible step would be to investigate these other dependencies, with respect to both the adult knowledge state and the available input. We have begun this process by testing relative clause formation and island constraints using the experimental design presented above. As predicted by syntactic theories, we found a subject island effect nearly identical to the one found above for wh-movement (we also tested several other island types with similar results). In the current project, we propose to experimentally investigate other displacement dependencies, and then continue the corpus annotation project to compare the relative frequencies of the aggregated dependency data with the acceptability results.

Second, linguistic theory concerning islands may provide natural sets of data to be included in the child's data intake, in addition to the subject and object-extraction data in our pilot study. In the decades since Ross's influential dissertation (1967), analyses have divided the various species of islands into genera along differing dimensions. For example, the weak versus strong distinction divides wh-islands, if/whether islands, negative islands, and factive islands from other island types based on the fact that d-linked wh-arguments (such as *which man*) can be moved out of the former but not the latter (Pesetsky 1987, Szalbolcsi 1990, see Szalbolcsi (2004) for a review); the *Condition on Extraction Domains* of Huang (1982) divides subject islands, sentential subject islands, and adjunct islands from the rest by proposing that non-complements such as subjects and adjuncts are islands because they are not governed. Furthermore, recent analyses of islands have also begun dividing island species into prototypical islands on one hand, and islands that require additional theoretical machinery on the other. For example, the barriers/phases approach of Chomsky (1986/2000) assumes that wh-islands are the prototypical island as manifested by the mechanics of the movement operation, with all other islands requiring additional theoretical machinery; the linearization approach of Uriagereka (1999) assumes that subject islands are the prototypical island as manifested by the linearization algorithm. For the purposes of this project, each of these analyses presents a different claim as to what types of information may be relevant for the learner. For instance, the strong/weak distinction can be interpreted as a claim that evidence relevant to wh-islands, negative islands, and factive islands may interact during the language learning process. So, the various theories of islands each make a claim as to what evidence should be considered relevant to the child. One of the primary responsibilities of Dr. Sprouse and one of the undergraduate research assistants will be to collect acceptability data on the structures relevant for each of these theories for comparison with their occurrence frequency in child-directed speech and adult conversational speech corpora.

Third, islands vary cross-linguistically and, in some cases, the variation appears to correlate with syntactic facts that may not seem immediately relevant to wh-dependencies. For example, it has been claimed that subject islands do not occur in languages that allow post-verbal subjects

such as Spanish and Italian (Torrego 1984, Uriagereka and Gallego 2007). If that is true, then children should be tracking the occurrence of post-verbal subjects in their language (and co-ocurrence with subject extraction) when learning if their language has subject islands. In this case, existing linguistic theory compels us to look for two facts. First, we would want to experimentally establish the existence or non-existence of subject islands in these languages. Second, we would want to investigate the correlation between post-verbal subjects and the non-existence of the subject islands constraint. Then, we can again adjust the data intake to our learning models accordingly, and see whether the induction problem still exists for subject islands. We have begun investigating these facts in Italian in collaboration with Ivano Caponigro at University of California, San Diego and Carlo Checcheto at the University of Milan, Bicocca. In our first experiment we examined several island constraints including subject islands with pre-verbal subjects, and found results identical to those in English: with pre-verbal subjects, there is a subject island effect in Italian. We are in the process of running a follow-up study to determine whether the island effect disappears with post-verbal subjects in Italian, as predicted by syntactic theory. We would also like to run similar experiments in other languages with post-verbal subjects such as Spanish.

As this brief discussion of subject islands has demonstrated, we have already begun collecting occurrence frequencies and acceptability judgments for island constraints in English using formal syntactic theories as a guide for determining which data may be relevant to the learning process. However, given the recent advances made by probabilistic models that incorporate more data than would traditionally be considered relevant (such as Perfors, Tenenbaum, & Regier (2006)), we wish to seriously consider the impact of indirect syntactic evidence. Much work is left to do, from investigating displacement dependencies other than wh-dependencies, to investigating the cross-linguistic variation of island constraints. These investigations will constitute the bulk of the empirical work of this project, forming the basis for the Bayesian models we will construct to test the UG hypothesis.

6. Roles of the principal investigators

The proposal as outlined above naturally divides itself between the PIs' respective areas of expertise. Dr. Sprouse brings knowledge of theoretical syntax, structural properties of language, and the empirical assessment of the adult knowledge state through experimental syntax. Dr. Pearl brings knowledge of language acquisition, corpus annotation and analysis, and computational learning models. Both Sprouse and Pearl will be involved with the annotation of the child-directed speech corpora and the adult conversational speech corpora, as well as supervising the undergraduate research assistants who will aid with this annotation. Sprouse will assess the adult knowledge state for the linguistic phenomena of interest, and supervise the assisting undergraduate research assistants. Pearl will assess the available input for the relevant linguistic phenomena in both corpora types. Sprouse will then identify possible learning biases based on current linguistic theory, and Pearl will evaluate the ability of both unbiased and biased learning models to reach the target state given the available input. Both Sprouse and Pearl will then analyze the nature of any biases found to be necessary, and determine which (if any) are likely part of UG.

7. Broader Impacts

With the recent advances in experimental syntax (Bard et al 1996, Keller 2000, 2003, Featherston 2004, 2005, Sprouse 2007, 2008), the availability of relevant electronic corpora (CHILDES: MacWhinney (2000); TalkBank), and the application of Bayesian learning models to language (Regier & Gahl 2004, Perfors, Tenenbaum, & Regier 2006, Pearl & Weinberg 2007, Foraker et al. 2007, Pearl & Lidz submitted), it is now possible to combine insights from theoretical linguistics, experimentation, and computational learning to begin seriously addressing the long-standing debate surrounding UG. The collaboration of Sprouse and Pearl will enhance

the infrastructure for research in this area by demonstrating how theoretical linguists and computational modelers can benefit from each other, and in turn, benefit the linguistics field at large. We also hope this collaboration will demonstrate how the concerns of nativists and non-nativists can be addressed simultaneously through the integration of multiple methodologies (e.g., experimental and computational).

In addition to providing a concrete way to clarify the UG hypothesis, this proposal will integrate research and education by furnishing invaluable hands-on research experience impacting theoretical linguistics to undergraduate students each year. While it is fairly common to find training in traditional psycholinguistics and computational linguistics, it is rare to find opportunities that deal with issues in theoretical linguistics. Due to the recent advances mentioned above in experimental and computational work impacting theoretical linguistics, these skills will not only be useful for the current research proposal but also for future lines of research surrounding the UG debate. Moreover, this proposal will enhance these students' discovery and understanding by giving them an enriched academic experience and continued access to professors engaged in highly relevant research that promises to have a significant impact on the field.

We will also be disseminating the results of this project in multiple ways. First, the annotated corpora we create will be made available through either CHILDES, TalkBank, or the Linguistic Data Consortium so that other researchers interested in complex linguistic phenomena can use these data. As mentioned in section 2, computational modeling work in these areas is often hindered by the lack of available corpora annotated appropriately. Our annotated corpora will thus provide a valuable resource for the community. Second, to stimulate activity across the field in this area, we propose to run two workshops at the University of California at Irvine (UCI), which is centrally located near other large research institutions with psychology and linguistics departments, such as the University of Southern California, the University of California at Los Angeles, and the University of California at San Diego. The first workshop will take place at the start of the project, near the end of the first year. It will involve 6-8 invitees as well as the PIs and interested students from UCI and surrounding universities. The initial workshop will have as a main objective identifying insights and issues from different approaches to tackling the language learning problem, including theoretical, experimental, and computational. The PIs would then explore the possibility of publication as a book or special issue of a journal. In addition, a workshop would be run at the end of the third year to publicize the PIs' own work and compare it to that of others. This will be of the same format as the initial workshop, with 6-8 invitees, the PIs, and interested students from UCI and surrounding universities. Potential invitees for both workshops include: Stephen Crain (theoretical, experimental), Jeffrey Elman (computational), Janet Fodor (computational), Adele Goldberg (theoretical, experimental), Sharon Goldwater (computational), Tom Griffiths (computational), Nina Hyams (experimental, theoretical), Mark Johnson (computational), Charles Kemp (computational), Howard Lasnik (theoretical), Jeffrey Lidz (theoretical, experimental), David Lightfoot (theoretical), Diane Lillo-Martin (experimental, theoretical), Brian MacWhinney (theoretical, computational), Elissa Newport (experimental), Amy Perfors (computational), Colin Phillips (theoretical, experimental), Terry Regier (computational), William Sakas (computational), Daniel Swingley (experimental, computational), Joshua Tenenbaum (computational), Michael Tomasello (experimental), and Charles Yang (computational). We also believe that during the first and subsequent years of the project, we will be in a position to present the ongoing results at major conferences both within the field of linguistics and across the broader field of cognitive science, including BUCLD (language acquisition), WCCFL (linguistics), GLOW (linguistics), and Cognitive Science (cognitive science). In addition, we expect individually and jointly to be able to submit manuscripts to leading journals, including *Cognition, Cognitive Science, Language Learning & Development,* Syntax, and *Linguistic Inquiry.*

As highlighted throughout this proposal, the research project envisioned here is potentially appealing to both the theoretical linguistics community and the broader linguistics community since it defines a methodology for addressing one of the most central debates surrounding language and learning - that of UG. Moreover, this project will demonstrate how this methodology can be applied to complex linguistic phenomena central to linguistic theory, such as islands. From the perspective of linguistic theory, the proposed research will be of interest to theoretical researchers of many persuasions, as it provides a method for assessing the learnability of whatever knowledge is deemed necessary for language and concretely integrates formal linguistic theory with empirical data. From the perspective of language acquisition, this research will offer a way to test the existence of induction problems for language learning and define what is necessary to solve those that do exist.

We are hopeful that the proposed research project will have an important effect on the methodological norms in both theoretical syntax and computational models of language acquisition. The experimental tools (such as magnitude estimation) for extracting quantifiable effects on judgment tasks from a large body of native speakers have not yet been widely implemented by theoretical researchers. Theoretical and psychological constraints on computational models of language learning (e.g., learning biases from linguistic theory, psychological plausibility considerations) have also not been widely used in tandem by computational researchers. The project we propose involves both of these extensively, and success with this approach would demonstrate their viability and potential importance to the field. We also expect this project to encourage theoretical and experimental linguists to join with computational linguists, and vice versa, when seeking answers to questions involving language and learnability. Sprouse and Pearl, as mentioned earlier, will provide an example of the benefits of this kind of collaboration.

In sum, we see the impact of our cross-disciplinary work on UG coming from two main factors. First, it concerns a central debate in linguistics, providing a defined methodology for answering questions about induction problems faced by language learners, learning biases that learners must possess, and the nature of those learning biases. Second, it connects theoretical and experimental work in linguistics with computational modeling, thus yielding results not achievable from these sub-fields individually. As this research refines our understanding of language learning and necessarily innate knowledge, we hope to inspire productive collaboration across the traditional divides of theory, experiment, and computation, both within UCI and across linguistics in general.