ABSTRACT

Title of Document:	NECESSARY BIAS IN NATURAL LANGUAGE LEARNING
	Lisa Sue Pearl, Doctor of Philosophy, 2007
Directed By:	Associate Professor, Department of Linguistics/UMIACS Associate Director, Neural and Cognitive Science Program Co-Director, Laboratory for Language and Media Processing Amy Weinberg

This dissertation investigates the mechanism of language acquisition given the boundary conditions provided by linguistic representation and the time course of acquisition. Exploration of the mechanism is vital once we consider the complexity of the system to be learned and the non-transparent relationship between the observable data and the underlying system. It is not enough to restrict the potential systems the learner could acquire, which can be done by defining a finite set of parameters the learner must set. Even supposing that the system is defined by *n* binary parameters, we must *still* explain how the learner converges on the correct system(s) out of the possible 2^n systems, using data that is often highly ambiguous and exception-filled. The main discovery from the case studies presented here is that

learners can in fact succeed provided they are biased to only use a subset of the available input that is perceived as a cleaner representation of the underlying system.

The case studies are embedded in a framework that conceptualizes language learning as three separable components, assuming that learning is the process of selecting the best-fit option given the available data. These components are (1) a defined hypothesis space, (2) a definition of the data used for learning (data intake), and (3) an algorithm that updates the learner's belief in the available hypotheses, based on data intake. One benefit of this framework is that components can be investigated individually. Moreover, defining the learning components in this somewhat abstract manner allows us to apply the framework to a range of language learning problems and linguistics domains. In addition, we can combine discrete linguistic representations with probabilistic methods and so account for the gradualness and variation in learning that human children display.

The tool of exploration for these case studies is computational modeling, which proves itself very useful in addressing the feasibility, sufficiency, and necessity of data intake filtering since these questions would be very difficult to address with traditional experimental techniques. In addition, the results of computational modeling can generate predictions that can then be tested experimentally.

NECESSARY BIAS IN NATURAL LANGUAGE LEARNING

By

Lisa Sue Pearl

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Amy Weinberg, Chair
Professor Jeffrey Lidz
Professor William Idsardi
Professor Charles Yang
Professor James Reggia

Dedications

To Amy Weinberg, who put up with a ridiculous number of last minute questions and drafts of *everything*. With good humor, to boot.
To Norbert Hornstein, who was always exceptional at convincing me that my ideas were any good. He's really quite persuasive.
To Charles Yang, whose work inspired all of this. More than once.
To Jeff Lidz, who helped me figure out how to write things that were actually comprehensible and who was always full of even more ideas.
To Bill Idsardi, who was also always full of even more ideas and who made me think very carefully about all my mathematical work.
To Philip Resnik, whose boundless energy and enthusiasm and questions completely inspired me time and time again.
To David Lightfoot, who taught the first theoretical linguistics class I ever took and who I wanted to grow up to be.
To Michelle Hugue, who taught me how to make computation into a tool and about how there are so many different ways to measure success.
To Peggy Antonisse, whose thoroughly sensible advice and calming influence were welcome balm in times of great stress and panic.

Acknowledgements

This work would have been a long time in coming without the listening ears, unwavering support, and good advice of many, many people, among them: Amy Weinberg, Norbert Hornstein, Charles Yang, Jeff Lidz, Bill Idsardi, Philip Resnik, David Lightfoot, Michelle Hugue, Peggy Antonisse, Colin Phillips, Andrea Zukowski, Heather Taylor, Brian Dillon, and Ivano Caponigro. This work was additionally supported by a National Science Foundation Graduate Research Fellowship.

Table of Contents

List of Tables

List of Figures

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) are easier for the learner to extract the correct systematicity from. Thus, even though

[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.

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.

Chapter 2: Bayesian Updating in a Linguistic Framework

The formal characterization of language learning from Yang (2002) consists of a language learning algorithm L, a set S of potential states the learner can be in, and experience from the linguistic environment E. The language learning algorithm L contains specifications for (a) the data intake the learner uses to update beliefs in available hypotheses and (b) the update procedure itself. In this chapter, I will describe the instantiation of the update procedure I will use for the case studies in the following chapters: an adapted form of Bayesian updating. Specifically, I will demonstrate how a standard implementation of this updating procedure (Manning & Schütze, 1999) can be adapted to language learning problems.

2.1 Bayesian Updating: Overview

Bayesian updating is a probabilistic updating procedure that is widely used in natural language processing tasks to update the probabilities of alternate available hypotheses (Manning & Schütze, 1999). Specifically, it calculates the conditional probability of the hypothesis, given the data. Probabilistic reasoning has been shown to be the optimal strategy for solving problems and making decisions given noisy or incomplete information (J. Pearl, 1996). Like many other systems, the linguistic system is often learned from observable data that is highly ambiguous and exception-filled. Thus, a probabilistic component seems necessary to the language learning mechanism.

There is also evidence for the psychological validity of a procedure like Bayesian updating as a method used by adult humans (Tenenbaum & Griffiths, 2001; Cosmides & Tooby, 1996; Staddon, 1988) and infants (Gerken, 2006). Specifically, these studies demonstrate probabilistic convergence on the more restrictive hypothesis compatible with the observable data. This is in line with the Bayesian updating procedure adopted here when there are two hypotheses under consideration that differ in their level of restrictiveness (section 2.1.5).

The main purpose of Bayesian updating is to infer the likelihood of a given hypothesis, given a series of examples as input. The implementation of Bayesian updating depends greatly on the structure of the hypothesis space, since the relation of the hypotheses to each other affects how probability is shifted between the different hypotheses. I will now examine several instances of hypothesis spaces below and their effect on Bayesian updating.

2.1.1 A Simple Case: Two Non-overlapping Hypotheses, Equally Likely

Suppose there are two non-overlapping hypotheses in the set: A and B. By non-overlapping, I mean that the examples in the input will either favor A or favor B unambiguously. There are no examples that signal (or can be accounted for by) both A and B – each hypothesis covers a distinct set of data points. Suppose also that the learner who will be using Bayesian updating has no reason to be biased towards one hypothesis, so the initial probabilities assigned to both A and B are 0.5. These are the prior probabilities associated with each hypothesis.

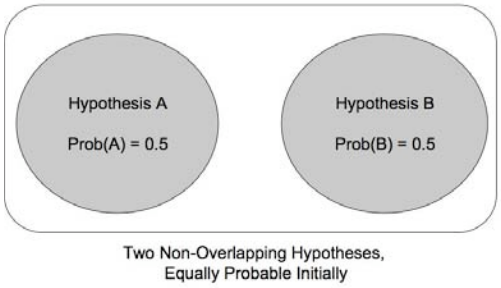
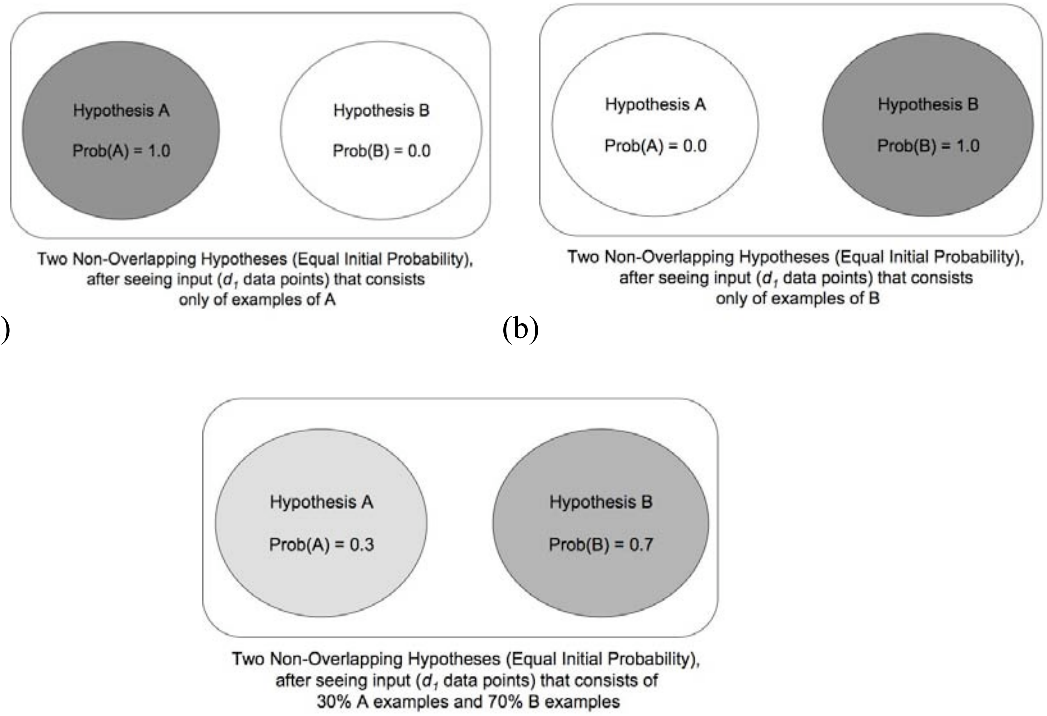


Figure 1. Two non-overlapping hypotheses, equally probable initially. The shading reflects how much probability is associated with each hypothesis.

The learner then encounters some amount of data (say d_1 data points) and uses Bayesian updating to shift the probability mass between A and B to reflect the distribution in the data intake. Each data point will cause the learner to shift the probabilities a small amount until the probability distribution among the hypotheses eventually matches the probability distribution encountered in the intake.


Hypothesis A
Prob(A) = 1.0
Hypothesis B
Prob(B) = 0.0
Two Non-Overlapping Hypotheses (Equal Initial Probability), after seeing input (d_1 data points) that consists only of examples of A
Hypothesis A
Prob(A) = 0.0
Hypothesis B
Prob(B) = 1.0
Two Non-Overlapping Hypotheses (Equal Initial Probability), after seeing input (d_1 data points) that consists only of examples of B
Hypothesis A
Prob(A) = 0.3
Hypothesis B
Prob(B) = 0.7
Two Non-Overlapping Hypotheses (Equal Initial Probability), after seeing input (d_1 data points) that consists of 30% A examples and 70% B examples

(a) (b)

(c)

Figure 2. Two non-overlapping hypotheses with equal initial probability after seeing various distributions of intake (the total amount is quantified as d_1 data points). The shading reflects how much probability is associated with each hypothesis.

If the data intake consists only of examples of A, the learner will eventually shift the probability so A is 1.0 and B is 0.0 (2a).[2] Conversely, if the data intake consists only of examples of B, the learner will eventually shift the probability so A is 0.0 and B is 1.0 (2b). In each of these cases, the learner shifts all the probability to a single hypothesis, thereby converging on one hypothesis as correct. However, it is possible that the learner will encounter a mixed distribution between A and B in the data intake. If so, the learner will shift the probability to reflect the bias in the perceived distribution since the target state is a probabilistic distribution between A and B. As a concrete example, if the input is consistently 30% A examples and 70% B examples, the learner will eventually shift the probability of A to be significantly less than that of B, reflecting the 30-70 distribution (2c).

2.1.2. A Variant on the Simple Case: Two Non-overlapping Hypotheses, with an Initial Bias for One Hypothesis

Suppose the hypothesis space again has two non-overlapping hypotheses, A and B. However, suppose the learner is biased towards A initially, so A has a higher prior probability associated with it than B does. For example, let the initial probability assigned to A be 0.7, and the initial probability assigned to B be 0.3. This scenario could represent a case where A is the default hypothesis and B is the exceptional (or marked) hypothesis – thus, B has a lower prior probability.

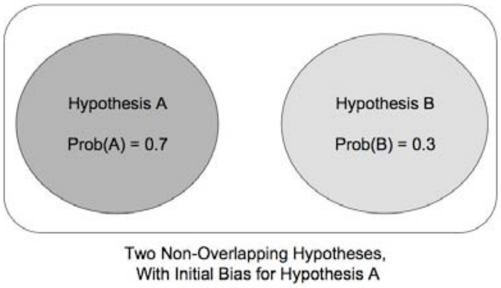
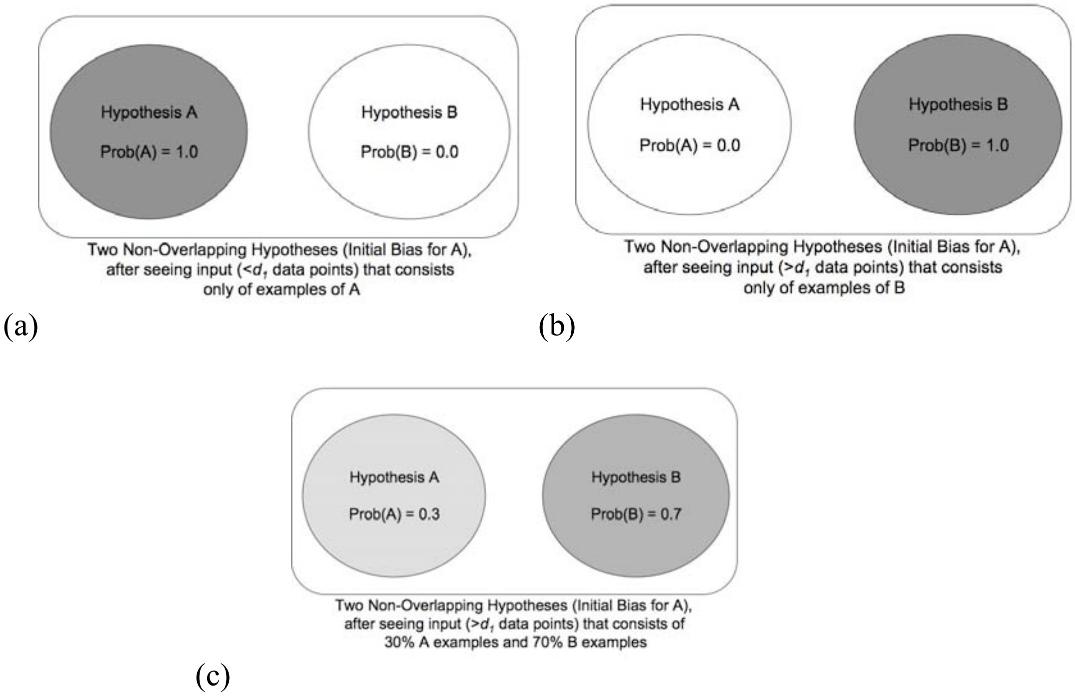

Hypothesis A
Prob(A) = 0.7
Hypothesis B
Prob(B) = 0.3
Two Non-Overlapping Hypotheses,
With Initial Bias for Hypothesis A

Figure 3. Two non-overlapping hypotheses, with an initial bias towards hypothesis A. The shading reflects how much probability is associated with each hypothesis.

The learner then encounters some amount of data and uses Bayesian updating to shift the probability mass between A and B to reflect the distribution in the data intake. As before, a learner encountering all A or all B examples will eventually shift the probability so that one hypothesis is 1.0 while the other is 0.0. However, because the prior probability of A is higher than that of B, it will take a smaller number of A examples to cause the probability of A to reach 1.0 (less than the d_1 data points in the unbiased hypothesis space) (4a). Conversely, since B is the disfavored hypothesis

[2] However, it is possible that the endpoints (0.0 and 1.0) will only be reached in the limit. Still, after encountering overwhelming data in support of one hypothesis over the other, the learner using Bayesian updating will likely be very *near* the endpoints. This point will hold true for all Bayesian updating examples in the remaining sections of this chapter.

initially, it will take a larger number of B examples to cause the probability of B to reach 1.0 (more than the d_1 data points in the unbiased hypothesis space) (4b). If the data intake has a mixed distribution, the same logic applies: a data distribution favoring A will be reflected more quickly in the probabilities the learner assigns to the hypotheses than a data distribution favoring B (4c).



Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input ($<d_1$ data points) that consists only of examples of A

(a)

Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input ($>d_1$ data points) that consists only of examples of B

(b)

Two Non-Overlapping Hypotheses (Initial Bias for A), after seeing input ($>d_1$ data points) that consists of 30% A examples and 70% B examples

(c)

Figure 4. Two non-overlapping hypotheses with an initial bias for hypothesis A after seeing various distributions and quantities of intake. The shading reflects how much probability is associated with each hypothesis.

2.1.3 A Less Simple Case: Two Overlapping Hypotheses, Equally Likely

Suppose there are two overlapping hypotheses in the set: A and B. By overlapping, I mean that there are two types of examples, unambiguous and ambiguous. Unambiguous examples either signal A or signal B. Ambiguous examples can be accounted for by both hypotheses. Thus, while each hypothesis has a unique subset of examples associated with it, there is also a subset that can be covered by both hypotheses. Suppose also that the learner has no reason to be biased towards one hypothesis, so the initial probabilities assigned to both A and B are 0.5.

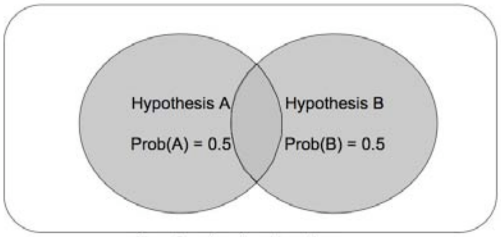


Figure 5. Two overlapping hypotheses, with equal probability initially. The shading reflects how much probability is associated with each hypothesis.

The learner then encounters some amount of data and uses Bayesian updating to shift the probability mass between A and B to reflect the distribution in the data intake. The important consideration is whether a given data point is unambiguous or ambiguous. If unambiguous (for either A or B), the updating will work the same as in the simple non-overlapping case, and the probability will be shifted slightly in favor of the hypothesis the data point is unambiguous for.

However, if the data point is ambiguous, the learning procedure must decide what to do with it. One possibility is to simply ignore the data point – this is the same as applying an unambiguous data filter that updates based only on unambiguous data points. This is a filter that will be explored in detail in chapters 4 and 5. Another possibility is to employ some strategy to deal with the ambiguous data point: use knowledge of the hypothesis space layout to assign partial credit (an approach explored in section 2.1.5 and chapter 3), use an informed guessing strategy (Fodor & Sakas, 2001), or randomly assign the data point to one hypothesis based on the current probabilities of both hypotheses (Yang, 2002). The random assignment method assumes that the effect of such ambiguous data will wash out in the face of the unambiguous data.

If the learner uses some strategy to extract information from an ambiguous data point in the overlapping hypothesis scenario, the learner will need to encounter more *total* data points than in the equivalent non-overlapping hypothesis scenario in order to converge on a hypothesis (more than d_l data points). This is simply a result of using both unambiguous and ambiguous data points to update the probabilities. Interestingly, if the learner uses an unambiguous data filter and ignores ambiguous data points, then we have a learning scenario that is very similar to the non-overlapping scenario: the learner must encounter d_l *unambiguous* data points in order to converge on the correct hypothesis. (In the non-overlapping hypothesis space, all data points are unambiguous.) Still, the total quantity of data points the learner encounters in the overlapping case will be greater than d_l, since the learner encounters both unambiguous and ambiguous data points. However, the only data points that cause any updating are the d_l unambiguous ones.

2.1.4 A Variant of the Less Simple Case: Two Overlapping Hypotheses, with an Initial Bias for One Hypothesis

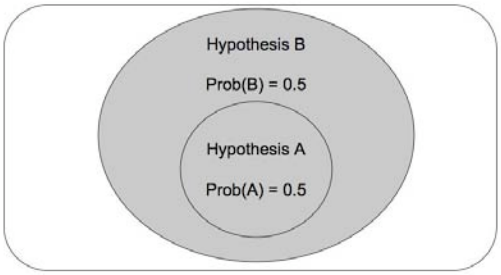
A variant of the overlapping case has biased initial probabilities. For instance, suppose hypothesis A has a prior probability of 0.7 while hypothesis B has a prior probability of 0.3. There are unambiguous examples of A, unambiguous examples of B, and ambiguous examples that can be accounted for by both A and B.

In terms of how the model deals with unambiguous and ambiguous data points, this scenario works the same as the unbiased overlapping scenario described in the previous section. The learner can either ignore the ambiguous data points, or employ some method to attribute them to one hypothesis.

However, as in the biased non-overlapping scenario described before, the number of data points the learner must encounter to converge on a hypothesis depends on how the data intake distribution relates to the prior probability distribution. If the data intake distribution is biased in the same direction as the prior probability distribution (say, 0.8 for A and 0.2 for B), the learner will need to encounter fewer data points to converge on the correct probability distribution. Conversely, if the data intake distribution is biased in the opposite direction from the prior probability distribution (say, 0.2 for A and 0.8 for B), the learner will need to encounter more data points to converge on the correct probability distribution.

2.1.5 An Even Less Simple Case: Two Overlapping Hypotheses in a Subset Relation, Equally Likely

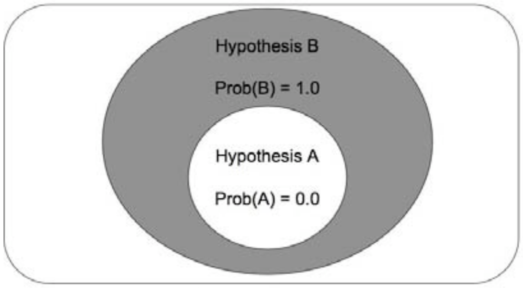
Suppose the hypothesis space again consist of two overlapping hypotheses, but one hypothesis is a subset of the other hypothesis. Let A be a subset of B, so all examples of A are also examples of B (Tenenbaum & Griffiths, 2001; Manzini & Wexler, 1987; Berwick, 1985; Berwick & Weinberg, 1984; Pinker, 1979). That is, while B has unambiguous examples, there are no unambiguous examples for A – all examples covered by hypothesis A can also be covered by hypothesis B. Suppose the initial probabilities assigned to both A and B are 0.5.



Two Overlapping Hypotheses in a Subset Relation,
Equally Probable Initially

Figure 6. Two overlapping hypotheses in a subset relation, with equal probability initially. The shading reflects how much probability is associated with each hypothesis.

Suppose the learner encounters only unambiguous examples for B in the data intake (say, d_2 data points). Eventually, the learner will shift all the probability to B (B = 1.0, A = 0.0).



Two Overlapping Hypotheses in a Subset Relation, after seeing input (d_2 data points) that consists only of examples of B

Figure 7. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing d_2 data points that are unambiguous for hypothesis B. The shading reflects how much probability is associated with each hypothesis.

But what if hypothesis A (the subset hypothesis) is the correct one for the target language? All examples covered by hypothesis A are also covered by hypothesis B – they are thus ambiguous data points. It is *impossible* for the learner to encounter any unambiguous data points for hypothesis A. If the data intake consists only of these ambiguous data points, one might expect the learner to remain at a neutral probability of 0.5 for each hypothesis since these data points are compatible with each hypothesis. The learner would be doomed never to converge on the correct hypothesis, the subset hypothesis A.

One way to save the learner from this fate is to exploit the layout of the hypothesis space. The Bayesian updating procedure can take advantage of the subset-superset relation of the hypotheses to favor hypothesis A when encountering an ambiguous data point. The logic is as follows:

(1) Logic of Favoring the Subset Hypothesis For an Ambiguous Data point

(a) If hypothesis B (the superset hypothesis) was correct, the data intake should contain at least *some* examples covered only in the superset B (i.e. unambiguous B examples).

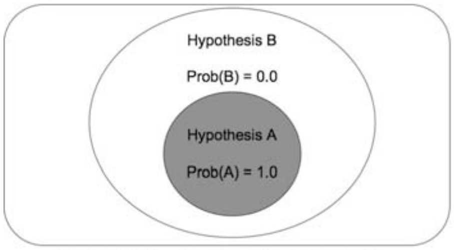
(b) If only examples covered by the subset A are encountered in the data intake, it becomes more and more unlikely that hypothesis B is correct.

(c) Therefore, the more the learner encounters only data points in the subset A (even though these are ambiguous data points), the more the learner will favor the subset hypothesis A.

A learner taking advantage of this logic will therefore consider a restriction to the subset A more and more probable as time goes on if only subset data points are encountered. This logic can be implemented in the Bayesian updating procedure itself, and has been referred to as the *size principle* (Tenenbaum & Griffiths, 2001).

Essentially, the smaller size of the set of examples covered by hypothesis A benefits hypothesis A when ambiguous examples are encountered. Specifically, the likelihood of encountering these examples given the smaller set covered by A is greater than the likelihood of encountering these examples given the larger set covered by B. So, A is slightly favored when encountering an ambiguous example covered in its subset.[3] After a sufficient number of ambiguous examples in the data intake (and, importantly for the basic version of the size principle, *no* unambiguous examples of the superset B), A will be highly favored.

We note that there is a disparity between the quantity of data points required to converge on B when using unambiguous data points as compared to the quantity required to converge on A using ambiguous data points. In particular, if the learner requires d_2 data points to reach probability p for B when encountering unambiguous B data points, the learner will require *more* than d_2 data points to reach p for A when encountering ambiguous data points. This is because the size principle allows A to only be *slightly* favored for an ambiguous data point while B is *exclusively* favored for an unambiguous B data point, though the actual amount of favoring depends on the relative sizes of A and B.

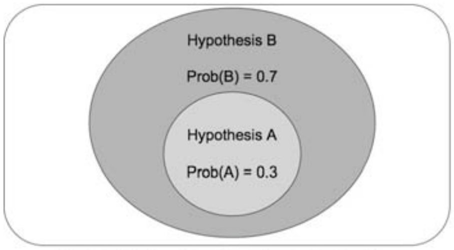


Two Overlapping Hypotheses in a Subset Relation, after seeing input (> d_2 data points) that consists only of examples of A

Figure 8. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing more than d_2 data points that are examples of A. The learner uses the size principle to converge on hypothesis A. The shading reflects how much probability is associated with each hypothesis.

If the data intake has a mixed distribution (both unambiguous B examples and ambiguous examples), the unambiguous B examples will have more effect on the learner's probability distribution than the ambiguous examples that slightly favor A. Both types of data points, however, will contribute to the final probability the learner converges on. Again, the number of data points required to converge on the final probability will be greater in this case (more than d_2 data points) than if only unambiguous B examples were encountered and the correct hypothesis was B exclusively.

[3] The amount A is favored depends on the relative sizes of A and B, which the learner must already know (perhaps as a separate prior) or empirically derive from the data. The smaller A is compared to B, the more A is favored given an ambiguous data point.



Two Overlapping Hypotheses in a Subset Relation, after seeing input (> d_2 data points) that consists of 30% A examples and 70% B examples

Figure 9. Two overlapping hypotheses in a subset relation with equal probability initially, after seeing more than d_2 data points that are a mix of unambiguous B examples and ambiguous examples in the subset A. The learner uses the size principle to converge on the probability that reflects the distribution observed in the input. The shading reflects how much probability is associated with each hypothesis.

It is important to note that exploiting the hypothesis space layout using the heuristic of the size principle is a non-trivial contribution to the learning problem for hypotheses arrayed in a subset-superset relationship. Though it is a heuristic and so not guaranteed to succeed for all cases, it nonetheless has an advantage over approaches that do not exploit the hypothesis space layout. Specifically, if only subset data are encountered, it will converge on the subset.

Suppose, however, that the learner did not use a heuristic like the size principle for learning. An instantiation of learning like this that still retains the advantages of probabilistic learning is the Naïve Parameter Learner (Yang, 2002), and the rate at which the learner shifts probabilities is represented by a parameter, gamma. A more conservative learner will have a smaller gamma, while a more liberal learner will have a larger gamma. For a data point, the Naïve Parameter Learner (NP learner) chooses one hypothesis and determines if the data point is compatible with it. If so, that hypothesis is rewarded while the remaining ones are punished; if not, it is punished while the remaining ones are rewarded. The update equations are given in (2), assuming two hypotheses, G1 and G2 (from Yang (2002)).

(2) Update equations for the NP learner for a hypothesis space with two hypotheses, G1 and G2, given a data point *d* and testing G1 against *d*

(a) If G1 is compatible with *d*,

$p_{G1} = p_{G1} + \text{gamma}*(1 - p_{G1})$

$p_{G2} = (1\text{-gamma})*p_{G2}$

(b) If G1 is not compatible with *d*,

$p_{G1} = (1\text{-gamma})* p_{G1}$

$p_{G2} = \text{gamma} + (1\text{-gamma})*p_{G2}$

To give a concrete example, suppose $p_{G1} = p_{G2} = 0.5$, and gamma = 0.005. Suppose data point *d* is encountered. The learner will test G1 with a 50% chance, and G2 with a 50% chance. Suppose the learner tests G1, and G1 is compatible with *d*. Then, the updated $p_{G1} = 0.5 + 0.005(1\text{-}0.5) = .5025$. The updated $p_{G2} = (1\text{-}0.005)*0.5 = 0.4975$.

As another example, suppose again that $p_{G1} = p_{G2} = 0.5$, and gamma = 0.005. Suppose data point *d* is encountered, and the learner tests G1 and finds it is not compatible with *d*. Then, the updated $p_{G1} = (1-0.005)*0.5 = 0.4975$, and the updated $p_{G2} = 0.005 + (1-0.005)*0.5 = 0.5025$.

As these two examples show for a hypothesis space that consists only of two hypotheses, when one hypothesis is punished by a certain amount, the other is rewarded by that same amount. If there were more than 2 hypotheses, the amount the tested hypothesis (G1) is punished/rewarded (gamma) would be distributed among the alternative hypotheses (G2…Gn).

The NP learner is implicitly driven by the availability of unambiguous data for one hypothesis – the alternative hypothesis is punished whenever it is used to interpret such unambiguous data points. Yet, if all data come from the subset hypothesis, then there will be no unambiguous data to punish the superset hypothesis. The NP learner encounters only ambiguous data, and is actually driven to convergence on *either* hypothesis, given sufficient data. This is shown in figure 10, assuming a hypothesis space where G1 is a subset of G2, and learning rates represented by gamma = 0.001 to 0.005, given 100,000 data points. The more liberal the learner is, the more likely the learner is to converge to one hypothesis or the other. Importantly, there is no guarantee that the learner will converge on the subset hypothesis, even though all data points come from the subset hypothesis.

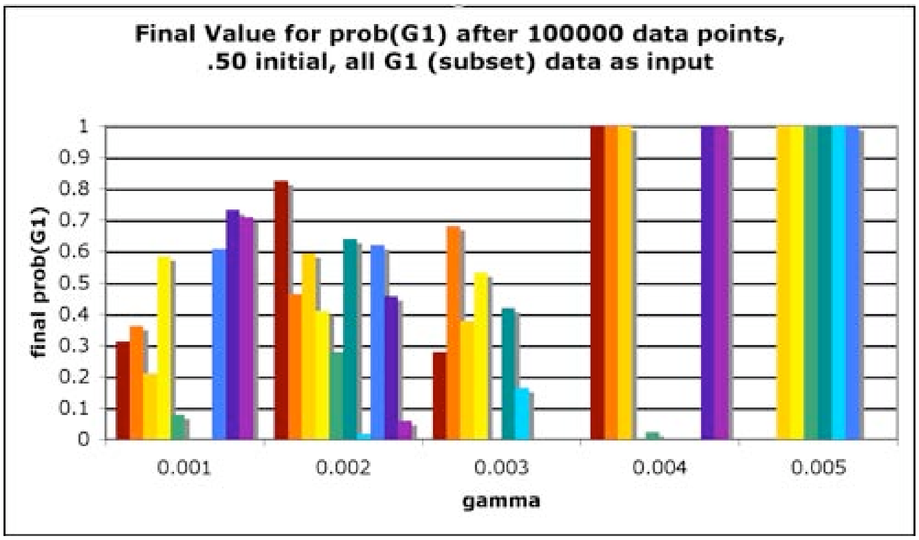


Figure 10. The NP learner, given ambiguous data from only the subset hypothesis, G1. This shows the results of 10 learners for each value of gamma, where gamma represents how conservative/liberal learning is. The NP learner has a tendency to converge to one hypothesis or the other, but is just as likely to converge to the subset G1 as the superset G2.

So, for learning cases where the hypotheses have a subset-superset relation to each other, approaches that do not exploit the hypothesis space layout will have difficulty converging on the subset hypothesis. The heuristic of the size principle provides a way to use this information to bias the learner towards the correct hypothesis.

2.1.6 Hypothesis Spaces for Language Learning

As we have seen, the layout of the hypothesis space and the relations between the hypotheses greatly affect how Bayesian updating uses the data intake to shift probability between alternate hypotheses. Crucially for Bayesian updating 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, the Bayesian updating procedure has nothing over which to operate. In short, if the learner has no options to select from, Bayesian updating cannot help. A Bayesian updating procedure dovetails with a defined hypothesis space; it does not replace it.

For language learning, a simple interpretation in the parametric framework of the generative tradition (Chomsky, 1981) is that there is a hypothesis space associated with each parameter, and alternative hypotheses within a given hypothesis space correspond to opposing values for linguistic parameters. For instance, suppose we examine the syntactic parameter of Verb-Second movement. A language with Verb-Second movement (such as German) will move the tensed Verb to the second phrasal position in the main clause; a language without Verb-Second movement (such as English) will not. The Verb-Second hypothesis space thus contains the hypotheses Verb-Second-Movement and No-Verb-Second-Movement. A learner of either German or English will encounter data points from the target language and use the data intake to converge on the appropriate hypothesis for that language.

In the remaining chapters, we will examine different hypothesis spaces in different domains of linguistics. Chapter 3 explores a language learning problem that spans syntax and semantics: English anaphoric *one*. Both the syntactic and semantic hypothesis spaces for English anaphoric *one* contain two overlapping hypotheses in a subset-superset relation, and these hypotheses are equally probable initially.[4]

Chapter 4 investigates a language learning problem in Old English syntax where the target state is a probabilistic distribution between two hypotheses, Object-Verb order and Verb-Object word order, that changes over time. The hypotheses are overlapping – that is, there are both unambiguous data points for each hypothesis and ambiguous data points. Both hypotheses are equally probably initially.

Chapter 5 studies the language learning problem of English metrical phonology, which is a data set plagued by noisy and contradictory data. There are nine separate interacting parameters, each with their own hypothesis spaces. Each hypothesis space contains two hypotheses that are overlapping, and these hypotheses are equally probable initially.

[4] Note that it is an assumption of the model that these hypotheses are equiprobable initially, rather than a derivation from theoretical work or an observation from experimental work.

2.2 Bayesian Updating: General Implementation for Language Learning in a Hypothesis Space with Two Hypotheses

I will now describe how the mathematical framework of Bayesian updating (Manning & Schütze, 1999) can be adapted to a language learning hypothesis space with two non-overlapping hypotheses, A and B.[5] The only data points a learner encounters will be unambiguous for either A or B. Note that we can use this same procedure for an overlapping hypothesis space (having both unambiguous data points and ambiguous data points) if the learner employs an unambiguous data filter that ignores the ambiguous data points. In this scenario, the only data points the learner uses to update the hypothesis probabilities are the unambiguous data points, which signal either A or B.

I will then briefly sketch how to modify the Bayesian update functions to account for an overlapping hypothesis space where the hypotheses are in a subset-superset relation. The details of this modification will be described more thoroughly in chapter 3, since the specific modifications are dependent on properties of the hypotheses themselves.

2.2.1 Updating with Unambiguous Data in a Hypothesis Space with Two Hypotheses

Suppose the hypothesis space consists of two hypotheses, A and B. Let the probability of hypothesis A be p_A and the probability of hypothesis B be p_B. Below, I describe how to update p_A. Before updating, p_A represents the prior probability of A; after updating, p_A represents the posterior probability of A. The calculation of p_B is straightforward once p_A is known, since $p_B = 1 – p_A$, given that there are only two hypotheses in the hypothesis space and only one of them can be correct for any given data point.

I assume that the learner extracts information only from the current data point, and uses the information from this data point to update the probabilities of the hypotheses. Thus, the sequence length for the language learning Bayesian update function is 1. Importantly, the learner does not store data points and subsequently conduct analyses across sequences of stored data points. So, the learner is not required to remember past data points in their raw form (i.e. as utterances), which I believe is a favorable quality for a model that aims to be psychologically realistic.

[5] Of course there are several alternative approaches for the updating procedure. For instance, one might try likelihood ratios (Neyman, J. & Pearson, E., 1928) to shift probability between hypotheses, given a data point. However, likelihood ratios require a prior knowledge of the success of the test used to identify the property of interest. Mapping this to the language learning problem, the learner would need to know the success of whatever method is used to identify unambiguous data for identifying *actual* unambiguous data. To know this, the learner must know what actual unambiguous data is. To know that, the learner would need to already know the system, so as to accurately determine what unambiguous data for it is. This, however, defeats needing to learn the system in the first place.

A more promising alternative is LaPlace's rule of succession (Manning & Schütze, 1999) which normalizes the number of previous successes (e.g. data points identified as unambiguous) against the total number of data points observed. Though similar to the adaptation of Bayesian updating used in this dissertation, it does not rely on a parameter corresponding to the period of fluctuation a learner is allowed. The benefit of this parameter (*t*) is discussed in section 2.3.3.

Because there are exactly two hypotheses in the hypothesis space, I use a binomial distribution to approximate a learner's expectation of the data distribution to be encountered. The binomial distribution is centered at p_A, so the learner's expectation is about the quantity of A data points that should be encountered in the data intake.

The binomial distribution is normally used to represent the likelihood of seeing *r* data points out of *t* total with some property. For example, if these are coin flip data points, the property might be "is heads". There are only two choices for each data point: the property is either present or absent. If these are coin flip data points, the coin is either heads or it isn't (specifically, it's tails). For the hypothesis space we are considering, the data point is either an example of A, or it isn't (specifically, it's an example of B). The highest confidence is assigned to the distribution where *r* A data points are observed our of *t* total: $r = t*p_A$. Recall that the binomial distribution is centered at p_A, and so the learner is most confident that the probability of seeing an A data point is p_A. So, *r* is the most probable number of A data points expected out of *t* total, given the current probability of hypothesis A, p_A.

As an example, suppose p_A is 0.5, as it is in the initial state in an unbiased hypothesis space before the learner has encountered any data points. The binomial distribution is centered at 0.5, which we can interpret as the learner having the most confidence that half the total data points encountered will be A data points. Specifically, the learner will expect $r = t*0.5$ data points to be A data points.

To update p_A after seeing a single unambiguous A data point *a*, we can follow Manning & Schütze's (1999) Bayesian updating algorithm and calculate the maximum of the *a posteriori* (MAP) probability. The a posteriori probability is the probability that p_A is the correct probability to center the binomial distribution at after seeing an unambiguous data point A; p_A represents the expected probability of encountering an A data point. We maximize this probability because we are using a probability distribution (specifically, the binomial distribution) to approximate the learner's expectation about the data distribution to be encountered. We want the maximum a posteriori probability that comes from using this probability distribution.

We represent the a posteriori probability as $\text{Prob}(p_A| \, a)^6$, and calculate it using Bayes' rule:

$$(3)\ \text{Prob}(p_A \,|\, a) \;=\; \frac{\text{Prob}(a \,|\, p_A) \;*\; \text{Prob}(p_A)}{\text{Prob}(a)}$$

We can now examine individual pieces of the right hand side equation. $\text{Prob}(a \mid p_A)$ is the probability of encountering the unambiguous A data point *a*, given that p_A is the correct probability to center the binomial distribution at. For a single instance (i.e. for the single data point *a*), the probability of encountering 1 instance of *a* for 1 observation from the binomial distribution centered at p_A is

[6] $\text{Prob}(p_A| \, a)$ is actually intended, rather than $\text{Prob}(A| \, a)$. This is because we are attempting to calculate the probability that p_A is the correct probability to center the binomial distribution at, given data point *a*. So, $\text{Prob}(p_A \mid a)$ can be thought of as shorthand for Prob(p_A is the correct center for binomial distribution that will match the distribution in the learner's intake | *a*).

$\binom{1}{1} * \mathrm{p_A}^1 * (1 - \mathrm{p_A})^{1-1}$, which is $\mathrm{p_A}$.

Prob($\mathrm{p_A}$) is the probability that $\mathrm{p_A}$ is the correct probability to center the binomial distribution at, i.e. that the learner should be most confident that an A data point will be encountered with probability $\mathrm{p_A}$. Recall that a binomial distribution centered at $\mathrm{p_A}$ will assign the highest confidence to the situation where $r = (\mathrm{p_A} * t)$ A data points are encountered out of t total. We can instantiate Prob($\mathrm{p_A}$) as the probability of encountering r A data points out of t total in a binomial distribution for *all* values of r, from 0 to t.[7]

$$(4)\ \mathrm{Prob}(\mathrm{p_A}) = \binom{t}{r} * \mathrm{p_A}^r * (1 - \mathrm{p_A})^{t-r} \text{ (for each } r,\ 0 \leq r \leq t)$$

Substituting these pieces back into equation (3) for the a posteriori probability yields (5):

$$(5)\ \mathrm{Prob}(\mathrm{p_A} \mid a) = \frac{\mathrm{p_A} * \binom{t}{r} * \mathrm{p_A}^r * (1 - \mathrm{p_A})^{t-r}}{\mathrm{Prob}(a)} \text{ (for each } r,\ 0 \leq r \leq t)$$

We can now calculate the MAP probability by finding the maximum of this equation. To do this, we take the derivative with respect to $\mathrm{p_A}$, set it equal to 0, and solve for $\mathrm{p_A}$.

(6) Calculating the MAP probability

$$\frac{d}{d\mathrm{p_A}}(\mathrm{Prob}(\mathrm{p_A} \mid a) = \frac{d}{d\mathrm{p_A}}\left(\frac{\mathrm{p_A} * \binom{t}{r} * \mathrm{p_A}^r * (1 - \mathrm{p_A})^{t-r}}{\mathrm{Prob}(a)}\right) = 0$$

$$\frac{d}{d\mathrm{p_A}}\left(\frac{\mathrm{p_A} * \binom{t}{r} * \mathrm{p_A}^r * (1 - \mathrm{p_A})^{t-r}}{\cancel{\mathrm{Prob}(a)}}\right) = 0 \text{ (since Prob}(a) \text{ is a constant w.r.t. } \mathrm{p_A})$$

$$\mathrm{p_A} = \frac{r+1}{t+1}$$

Recall that r is the previous expected number of A data points encountered out of t data points total. Hence, $\mathrm{r} = \mathrm{p_{A\,old}} * t$. Therefore, we write the update function for $\mathrm{p_A}$ after encountering unambiguous A data point a as (7a).

(7a) Update function for $\mathrm{p_A}$ after seeing unambiguous A data point a

$$\mathrm{p_A} = \frac{\mathrm{p_{A\,old}} * t + 1}{t+1}$$

[7] Note that approximating Prob($\mathrm{p_A}$) this way is a non-standard assumption. However, it yields update equations with psychologically desirable properties that other more standard assumptions do not.

An intuitive interpretation of this update function is that the numerator represents the learner's confidence that the encountered unambiguous A data point *a* is a result of the A hypothesis being correct; the denominator represents the total data encountered so far. Thus, 1 is added to the numerator because the learner is fully confident that the unambiguous data point *a* indicates the A hypothesis is correct; 1 is added to the denominator because a single data point has been encountered.

As we observed before, given that there are only two hypotheses in the hypothesis space, we can calculate the new p_B after seeing an unambiguous A data point *a* as $p_B = 1.0 - p_A$.

(7b) Update function for p_B after seeing unambiguous A data point *a*

$$p_B = 1 - p_A = 1 - \frac{p_{A\text{ old}} * t + 1}{t + 1}$$

Now, we can also derive the update functions for p_A and p_B after seeing an unambiguous B data point *b*. The derivation of the update function for p_B after seeing *b* is identical to the derivation of the update function for p_A after seeing *a*, and leads to equation (8).

(8) $$p_B = \frac{p_{B\text{ old}} * t + 1}{t + 1}$$

Again, since there are only two hypotheses in the hypothesis space, $p_B = 1.0 - p_A$. So, if we wish to track the value of p_A, we can substitute this into equation (7) and derive the update function for p_A after an unambiguous B data point *b* is encountered.

(9)

$$p_B = \frac{p_{B\text{ old}} * t + 1}{t + 1}$$

$$(1 - p_A) = \frac{(1 - p_{A\text{ old}}) * t + 1}{t + 1}$$

$$p_A = 1 - \frac{(1 - p_{A\text{ old}}) * t + 1}{t + 1} = \frac{t + 1 - (t - p_{A\text{ old}} * t + 1)}{t + 1}$$

$$p_A = \frac{p_{A\text{ old}} * t}{t + 1}$$

This update equation is identical to (7a), except that 0 is added to the numerator instead of 1. This reflects the intuitive notion that the learner should have no confidence that the A hypothesis generated the unambiguous B data point *b* just encountered.

2.2.2 Updating with Ambiguous Data in a Hypothesis Space with Two Hypotheses

We have just seen how to derive the update functions for when an unambiguous data point is encountered. Suppose, however, that the learner encounters an ambiguous data point and does not impose a filter that ignores such data for the purposes of updating. Since this data point is ambiguous between hypotheses A and B, the value added to the numerator should be a reflection of the learner's confidence that the data point indicates each of these hypotheses.

I now focus on the update of p_A (recalling, of course, that we can easily derive p_B as $1 - p_A$). If an unambiguous A data point is encountered, 1 is added to the numerator to indicate full confidence in A (and no confidence in B). Conversely, if an unambiguous B data point is encountered, 0 is added to the numerator to indicate no confidence in A (and full confidence in B). So, if a data point is ambiguous between the two hypotheses, a value greater than 0 and less than 1 should be added to the numerator. If the value added is 0.5, this would reflect no bias for either hypothesis (a truly ambiguous data point); if the value added is closer to 1, this would reflect a bias for the A hypothesis; if the value added is closer to 0, this would reflect a bias for the B hypothesis. As an example, if the learner has reason to favor A (perhaps because A and B are in a subset-superset relation with A as the subset), the value added would be greater than 0.5 but less than 1. The exact value would depend on the relative size of the sets of examples covered by A and B.

(10) Hypothetical update function for p_A after encountering an ambiguous data point, A is a subset of B, and so there is bias for hypothesis A

$$p_A = \frac{p_{A\ old} * t + m}{t + 1},\ 0.5 < m < 1$$

It is important to note that ignoring ambiguous data is *not* equivalent to adding 0.5 to the numerator when encountering an ambiguous data point. One might presume this since we interpreted the addition of 0.5 to the numerator as having no bias for either hypothesis. The crucial difference is in the invocation of the update function: if the ambiguous data point is ignored, no updating occurs; if the ambiguous data point is used, the update function is invoked. This has important consequences if the learner employs the strategy of adding 0.5 to the numerator when encountering an ambiguous data point. Each ambiguous data point will cause an update that will drive p_A closer to 0.5.

As an example, suppose the input stream contains 10% unambiguous A data points and 90% ambiguous data points. If the learner imposes an unambiguous data intake filter, the learner will only update p_A for 10% of the data points encountered and will always add 1 to the numerator. This results in a p_A that is significantly greater than 0.5 (though possibly still less than 1). Conversely, if the learner updates for both unambiguous and ambiguous data points, the update function is always invoked; 10% of the time, p_A is pushed closer to 1.0 but 90% of the time p_A is pushed back towards 0.5. This results in a p_A that is significantly closer to 0.5 than the p_A obtained by using only unambiguous data to update. In short, the learner is less likely to converge on the correct hypothesis, A.

2.2.3 About *t*

The update functions just derived depend on two parameters: the prior probability, $p_{A\ old}$, and the total amount of data expected during the learning period, *t*. Expecting the learner to already know the prior probability seems reasonable, as it is the most recent value the learner has calculated using the update function. Expecting the learner to already know the total amount of data during the learning period, however, may seem farfetched. Yet, the underlying concept behind *t* can also be interpreted as the amount of change a real learner's brain is allowed to undergo before settling into the final state. This would be a biologically given constraint. In my simulations, this amount is simply quantified as the total amount of data available as intake to the learner (i.e. the learner can use *t* data points of data to update the probabilities assigned to the different hypotheses).

The role of *t* in the update functions is to determine how much the probability should be shifted, given a single data point. If *t* is small, a single data point shifts the probability a great deal. This is a direct result of the fact that a small *t* means the expected data set will be small, and so only a small number of changes are allowed. Thus, the learner shifts the probability more liberally in an attempt to get to an appropriate target state before *t* runs out. Conversely, if *t* is large, a single data point shifts the probability a lesser amount. This is a direct result of the fact that a large *t* means the expected data set will be large, and so a large number of changes are allowed. The learner in this case can afford to be more conservative when shifting probability because there are more chances to shift the probability before *t* runs out.

Importantly (and perhaps surprisingly), the value of *t* is essentially arbitrary: the final probability the learner settles on is independent of the size of *t*, provided *t* is not *too* small. The reason for this stability is that the behavior of the learner is dependent on the probability distribution of the data. As long as *t* is large enough for the learner to observe a reasonably accurate sample of the probability distribution in the data intake, the learner will converge on a final probability that is the same across different values of *t*. If *t* is small, each data point has a larger impact; if *t* is large, each data point has a smaller impact. The final probability, however, does not change. This will be demonstrated with an explicit example in the chapter 3.[8]

The parameter *t* can also capture the notion of "critical period" or "period of fluctuation", where learning of particular aspects of the linguistic system ceases abruptly after some maturational point. Specifically, after the learner has encountered *t* amount of data in the intake, no more updating is possible. The probabilities for the

[8] Note that this is different from saying that that *t* must be empirically determined for each learning problem. It does not matter what *t* is for a given learning problem– the bias in the distribution is what drives the learner one way or the other. The value of *t* simply quantifies the amount a given data points alters the learner's associated probabilities for each hypothesis.

hypotheses are set, and future data points encountered have no effect. In short, the data intake for this hypothesis space is then zero, no matter what the available input is. This maps directly to the idea of a cut-off point for language learning, after which no further input can influence the learner's linguistic hypotheses.

Equipped with these relations between the period of fluctuation, *t*, and the data intake, I can speculate on the time course of parameter-setting for individual parameters. In this model, the period of fluctuation is defined by *t*: the size of *t* determines the length of the period of fluctuation. If we link *t* to the amount of change a real learner's brain is allowed to undergo and so view *t* as a biologically given constraint, we might expect that *t* should be invariant across different parameters. If all parameters have the same period of fluctuation (as defined by *t*), we should expect all parameters to be set at the same time. Yet, there is ample evidence that this is not the case. How do we reconcile this with our view of *t*?

The answer lies in the relation between *t*, data intake, and the filtering component of the learning theory. The period of fluctuation is defined by a constant value of *t*, but *t* is defined over the quantity of data points in the *intake* - not just in the available input. The proportion of input that is used as intake can vary from parameter to parameter, based on the filters used to define intake. High proportions of intake from input will allow the quantity of intake to accumulate more quickly over time; low proportions of intake from input will cause the quantity of intake to accumulate more slowly over time. The more quickly intake is accumulated over time, the faster the learner reaches the data intake limit of *t*. So, this view predicts that parameters that accumulate data intake more quickly will be set earlier than parameters that accumulate data intake more slowly.

As a concrete example, suppose learners implement an unambiguous data filter that causes the data intake to consist only of unambiguous data. The time course of parameter-setting should then depend on the quantity of unambiguous data available in the input. Yang (2004) provides a summary of evidence from experimental studies that suggests this is precisely what happens for certain syntactic parameters, including the information in table 2.1.[9] Syntactic parameters with a larger proportion of unambiguous data in the input are acquired earlier while syntactic parameters with a smaller proportion of unambiguous data in the input are acquired later.

[9] This is no longer true if the learner uses ambiguous data as well, unless the combination of unambiguous and ambiguous data used yields these same correlations. Again, one possibility is there is a correlation between the data intake (*useable* data) and the time course of acquisition. In that case, *t* would again represent the amount of change allowed, but useable ambiguous data "uses up" some of *t* (in addition to unambiguous data using up some of *t*). It is even possible that unambiguous data would use up more of *t* than useable ambiguous data would, perhaps in proportion to the amount of perceived ambiguity: the more unambiguous the data is, the more *t* is used up since the learner is more confident that the data is informative.

Parameter	Target Language	Unamb Data Frequency	Time of Acquisition
Verb-Raising[a]	French	7 %	1;8 (Pierce, 1992)
Obligatory Subject[b]	English	1.2%	3;0 (Valian, 1991)
Verb-Second[c]	German/Dutch	1.2%	3;0-3;2 (Clahsen, 1986)
Scope-Marking[d]	English	0.2%	4;0+ (Thornton & Crain, 1994)

Table 2.1. The effect of data intake accumulation on parameter-setting. Assuming an unambiguous data filter, syntactic parameters that have a higher proportion of input used as intake are the parameters that are acquired earlier.

Looking at the data in table 2.1, we can see the relation between the frequency of unambiguous data in the learner's input and the time of acquisition. We look first at Verb-Raising. In languages like French, the tensed verb moves before adverbs negation and adverbs ('*Jacques voit souvent/pas Simone*'; 'Jack sees often/not Simone'), in contrast to languages like English ('Jack often sees Simone') (1a). Unambiguous data signaling Verb-Raising comprise about 7% of the input, and children appear to have knowledge of Verb-Raising quite early.

We turn then to the Obligatory Subject. In languages like English, a subject is required ('He saw Rafael', 'It is raining'), while in languages like Spanish, the subject is optional ('*(Él) vio a Rafael', 'Llueve*'; '(He) saw Rafael', 'Rains'). Unambiguous data for Obligatory Subject is much less frequent than Verb-Raising data, and the time of acquisition is also later than that of Verb-Raising.

We can look to Verb-Second as well. In languages like German and Dutch, the tensed Verb in the main clause is moved to the second phrasal position, following one phrase of any type ('*Ich liebe die Katzen', 'Die Katzen liebe ich'; '*I-Subj love cats-Obj', 'Cats-Obj love I-Subj'). Unambiguous data for Verb-Second appear approximately as frequently as unambiguous data for Obligatory Subject, and the time of acquisition is also approximately equivalent.

There is also evidence from Scope-Marking. In German, Hindi, and other languages, long-distance *wh*-questions leave intermediate copies of *wh*-markers ('*Wer glaubst du wer Recht hat?'*, 'Who think you who right has?', Who do you think has the right?). For English children to know that English does not use this option, long distance *wh*-questions must be heard in the input, a type of data that is very infrequent in the available input to children. And indeed, the time of acquisition is much later.

In summary, children could learn from a fixed quantity of *relevant* data points, irrespective of parameter, and this would accord with experimental evidence. The quantity is constant across all parameters (*t*), but the availability of relevant data (intake) is not constant across all parameters. This yields different time courses of parameter-setting.

2.3 Summary of Bayesian Updating Adapted to a Linguistic Framework

I have now described the mathematical framework I will employ in the subsequent chapters to explore different case studies in language learning. In addition, I have sketched how values integral to the mathematical framework can be mapped to already existing concepts in the language learning literature. This framework will be the basis for the updating procedure used by the learner to shift probability between competing hypotheses. I reiterate that this updating procedure is domain-general, and is applicable across linguistic domains (and other cognitive domains). However, the representations assumed for the hypothesis space and the filters tested in each case study will be domain-specific. The separation of a learning theory into three distinct parts allows us to merge domain-specific components with domain-general components and thus have a theory that is both.

Chapter 3: The Case of Anaphoric *One*

3.1 Anaphoric One: The Necessity of Domain-Specific Constraints

The phenomenon under investigation is the interpretation of the anaphoric element *one* in English. Previous work has argued that infants' knowledge of anaphoric *one* could not be derived from their experience with this form (Lidz & Waxman, 2004; Lidz, Waxman, & Freedman, 2003). Instead, it was argued that the learner must be equipped with prior constraints on the hypothesis space. Because of these constraints, certain interpretations are simply never considered as potential hypotheses – specifically for anaphoric *one*, the learner would not consider the hypothesis that *one* is anaphoric to N^0. These constraints were described as being part of the domain-specific representational format for language learning. However, subsequent work (Regier & Gahl, 2004) replied that a probabilistic learner could acquire this knowledge using the domain-general learning procedure of Bayesian updating. No constraints on the hypothesis space (or domain-specific constraints of any other kind) would then be required. Regier and Gahl (henceforth R&G) provided their learning model with a small set of hypotheses to choose from that were derived from domain-specific representational content. Because there are no constraints on the hypothesis space, R&G's model considers more hypotheses than the learner of Lidz, Waxman, and Freedman (henceforth, LWF).

The two sides are then set up. The LWF learner requires a hypothesis space defined over domain-specific representations, as well as domain-specific constraints that preclude certain hypotheses from being considered. No filters on data intake are posited, and the learner is compatible with a Bayesian updating procedure. The R&G learner also requires a hypothesis space defined over domain-specific representations, but does not require additional constraints on the hypothesis space. Instead, the R&G learner rules out the incorrect hypotheses using a particular implementation of Bayesian updating that exploits the layout of the hypothesis space. R&G also do not *explicitly* posit filters on data intake, and thus claim that no additional information beyond probabilistic updating is required to converge on the correct interpretation of anaphoric *one.*

However, I will argue that R&G's conclusion was too quick. In particular, the R&G learner considers only a restricted source of evidence, which inflates the estimate of the learner's success. By restricting the data intake this way, this model in fact *implicitly* implements two domain-specific filters on the learner's data intake, which will be discussed in detail later in the chapter. However, when a model of a learner that is in the true spirit of the R&G proposal is set up, i.e. one that has no filters on data intake, we will find that this unconstrained Bayesian learning model does not display the correct behavior. If the learner considers the full array of evidence in the input, the learner will fail to learn the correct interpretation of anaphoric *one.*

A Bayesian model without domain-specific constraints is plagued by a particularly pernicious problem in language learning. Specifically, representations across domains are aligned (e.g. strings of words project to interpretations about referents in the world). In the case studied here, when we allow the learner to consider the correspondences across levels of representation (syntax and semantic reference), we find that an unrestricted Bayesian model fares very poorly. This conclusion casts doubt on Bayesian learning as the sole source of constraints on learners. In short, this case suggests that the overly general nature of domain-general learning must be reigned in by domain-specific representations and domain-specific filters on data intake.

3.2 Why Learning Anaphoric One Is Interesting

To learn the correct interpretation of anaphoric *one*, it is believed that the learner must consider both the syntactic level of representation and the semantic level of representation. At the syntactic level, the learner must learn what the linguistic antecedent of *one* is; at the semantic level, the learner must determine what object in the world the noun phrase 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.* A linguistic antecedent (syntax) can be translated into a reference to an object in the world (semantics) and so both syntactic and semantic representations are implicated in knowledge of *one*. As we will see below, the correct syntactic representation for English adults is that the linguistic antecedent of *one* is a string classified as the category N'. This syntactic knowledge has semantic consequences, which are what LWF used to determine if 18-month olds preferred that specific syntactic representation. In this way, we can see that the knowledge that *one* refers to N' strings traverses both the syntactic domain and the semantic domain.

Acquisition of anaphoric *one* is an interesting learning problem because the data that would lead a learner to the correct representation are quite sparse. In particular, LWF estimated that less than 0.3% of the child's input containing anaphoric *one* provided unambiguous evidence for the correct representation. Moreover, the rate of ungrammatical sentences containing anaphoric *one* was twice this amount, making the occurrence of useful (unambiguous) data below noise level. Given this pattern of data, LWF argued (following Baker (1979) and Hornstein & Lightfoot (1981)) that constraints on the representation of anaphoric *one* must be built into the learner's domain-specific representations. The learner should never consider hypotheses where *one* refers to categories smaller than N', such as N^0.

R&G countered that a learner using a domain-general Bayesian learning procedure could converge on this knowledge by using ambiguous data with certain properties. This particular class of ambiguous data functions as indirect negative evidence for the correct hypothesis[10]. Using this ambiguous data, they argued, would make the proposed constraint on the linguistic representations unnecessary. The appeal of a domain-general learning procedure without domain-specific filters resides in the lack of biases found inside the learner. However, R&G's model made use of only *some* of the available ambiguous data and of only *semantic* data to converge on the syntactic representation. This decision implements two domain-specific filters on

[10] But see Lasnik (1987) for comments about indirect negative evidence in language learning.

the learner's data intake. I will investigate the results of a probabilistic Bayesian learning procedure that removes these filters.

The procedure I develop uses all the available ambiguous data as well as both syntactic and semantic data to converge on the probabilities of competing representations. I will show that, even under the most generous estimates of the various parameters involved in such a model, a Bayesian learner lacking domain-specific filters on data intake will fail to converge on the syntactic knowledge that *one* is anaphoric to N' strings and fail to have the standard adult interpretation of what set of referents in the world *one* can refer to. In short, the unconstrained Bayesian learner will not learn the preferred adult interpretation of anaphoric *one*, in contrast to what real children do.

The chapter proceeds as follows. First, I will briefly describe the relevant parts of the grammar of anaphoric *one*. I will then review the behavioral evidence indicating 18-month olds have acquired the adult representation of anaphoric *one* and the argument by LWF that the input available to children is too sparse to support acquisition of this knowledge. Then, I address various proposals to circumvent the sparse data problem, and detail how the R&G proposal about a domain-general solution to this problem implicitly implements domain-specific filters on the data intake. Following that, I describe a Bayesian learning model that is truly domain-general, in that it removes all implicit filtering on the data. I show that such a model fails to acquire the adult interpretation of anaphoric *one*. In addition, I describe how under a less charitable assumption of a certain parameter value, the Bayesian learning model would perform even more poorly. Then, I identify the source of the model's failure. One contributing factor to the spectacular failure of the model derives from the link between syntax and semantics. A second contributor to this failure is the abundance of ambiguous data, which given Bayesian learning techniques, causes to the learner to misconverge. I argue that successful acquisition depends on a domain-specific filter on the data. Finally, I speculate on the origin of the necessary domain-specific filter, suggesting that its roots may lie in a syntactocentric approach to learning anaphoric *one*.

3.3 Anaphoric One

3.3.1 Adult Knowledge: Syntactic Categories and Semantic Referents

For English adults, the element *one* is anaphoric to strings that are classified as N' (i.e., the antecedent for *one* is an N' string), as in example (1) below. The structures for the N' strings are represented in figure 11.[11]

(1a) *One* is anaphoric to N' (*ball* is antecedent)
"Jack likes this *ball* and Lily likes that *one.*"
(1b) *One* is anaphoric to N' (*red ball* is antecedent)
"Jack likes this *red ball* and Lily likes that *one*."

[11] Note that the precise labels of the constituents here are immaterial. If the structure is [$_{DP}$ this [$_{NP}$ red [$_{NP}$ ball]]], the conclusions reached in this chapter would not be changed.

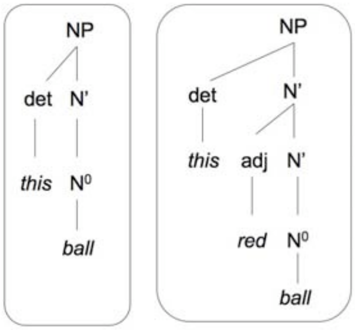


Figure 11. Structures for the N' strings *this ball* and *this red ball*.

These representations encode two kinds of information: constituency structure and category structure. The constituency structure tells us that in a Noun Phrase (NP) containing a determiner (det), adjective (adj) and noun (N^0), the adjective and noun form a unit within the larger Noun Phrase. The fact that *one* can be interpreted as a replacement for those two words (as in (1b)), tells us that those two words form a syntactic unit. The category structure tells us which pieces of phrase structure are of the same type. That is, both *ball* and *red ball* are of the type N'. The following argument explains how we can conclude this.

Consider the following examples in which *one* cannot be anaphoric to a noun (cf. Baker (1979)):

(2i) a. Jack met the member of Congress…
b. * …and Lily met the one of the Society for Creative Anachronism.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ member] [$_{PP}$ of Congress]]]

(2ii) a. Jack reached the conclusion that syntax is innate…
b. * …and Lily reached the one that learning is powerful.
c. [$_{NP}$ the [$_{N'}$ [$_{N0}$ conclusion] [$_{CP}$ that syntax is innate]]]

These contrast with cases in which what follows the head noun is an adjunct/modifier. Here, *one* can substitute for what appears to be only the head noun.

(2iii) a. Jack met the student from Peoria…
b. … and Lily met the one from Podunk.
c. [$_{NP}$ the [$_{N'}$ [$_{N'}$ [$_{N0}$ student]] [$_{PP}$ from Peoria]]]

(2iv) a. Jack met the student that Lily invited to the party
b. … and Lily met the one that Jack invited.
c. [$_{NP}$ the [$_{N'}$ [$_{N'}$ [$_{N0}$ student]] [$_{CP}$ that Lily invited to the party]]]

These cases differ with respect to the status of what follows the noun. In (2i) and (2ii) what follows the noun is a complement, but in (2iii) and (2iv) what follows the noun is a modifier. We can see that *one* can take a noun as its antecedent only when that noun does not take a complement. I will represent this by saying that *one* must take N' as its antecedent and that in cases in which there is no complement, the noun by itself is categorized as both N^0 and N’. In other words, in cases like (1a), it must be the case that *ball* = N’, as in the structure in Figure 11. If it weren’t, we would have no way to distinguish this case from one in which *one* cannot substitute for a single word, as in (2i) and (2ii).

3.3.2 The Pragmatics of Anaphoric *One*

In addition, when there is more than one N’ to choose from (as in (1b) above), adults generally prefer the N’ corresponding to the longer string (*red ball*). For example, in (1b) an adult (in the null context) would often assume that the ball Lily likes is red – that is, the referent of *one* is a ball that has the property red (cf. Akhtar et al. (2004)). This semantic consequence is the result of the syntactic preference for the larger N’ *red ball.* If the adult preferred the smaller N’ *ball*, the semantic consequence would be no preference for the referent of *one* to be red, but rather for it to have any property at all. Importantly, though, this preference is not categorical. It is straightforward to find cases where it is overridden, as in (3):

(3) Jack likes the yellow bean but Lily likes that one.

Here, it is quite easy to take *one* to refer to *bean* and not *yellow bean*.

3.3.3 Children’s Knowledge of Anaphoric *One*

But do children prefer *one* to be anaphoric to an N’ string (and more specifically the larger N’ string if there are two), rather than to an N^0 string? If so, the semantic consequence would be readily apparent: the antecedent for *one* would be phrasal, and hence the referent of *one* would be sensitive to properties mentioned by modifiers in the antecedent. LWF conducted an intermodal preferential looking paradigm experiment (Golinkoff et al., 1987; Spelke, 1979) to see if infants did, in fact, have a preference for the referent of *one* to have properties mentioned by the modifier in the antecedent (i.e., for a red bottle if a potential antecedent of *one* is *red bottle*).

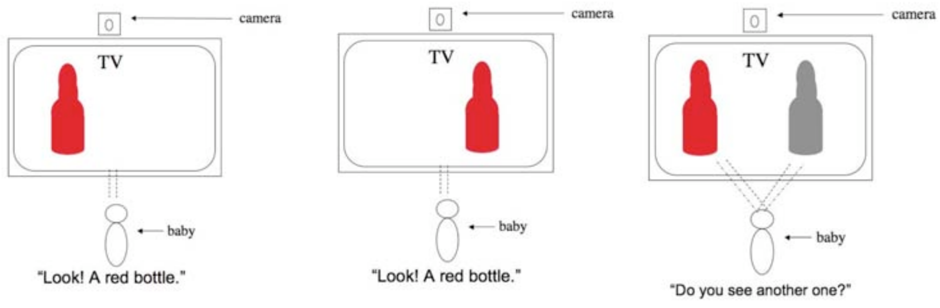


Figure 12. LWF experimental set up.

The infant in the LWF experiment is first shown a bottle of one color while several utterances of the form “Look! An *adjective* bottle.” are played simultaneously. Then, in the test stage, two bottles are shown – one of the *adjective* color and one of another color. The utterance “Do you see another one?” is played simultaneously and the infant’s looking preferences are recorded.

The 18-month olds demonstrated a significant preference for looking at the bottle that had the same property mentioned in the N’ string – e.g. the bottle that was red when the N’ string *red bottle* was a potential antecedent. These same results were obtained when the infants listened to, “Look! An *adjective* bottle” followed by “Do you see another *adjective* bottle?” (See Lidz & Waxman (in prep.) for more empirical data supporting this.) This suggests that the infants were interpreting these utterances similarly, namely that *one* referred to *adjective bottle* in the original test condition.

Notably, the infants’ response differed from the baseline condition where they heard, “Look! An *adjective* bottle” followed by “What do you see now?” In the baseline condition, the infants had a novelty preference and looked longer at the non-*adjective* bottle, e.g. a non-red bottle if they had previously seen a red bottle and heard, “Look! A red bottle”.

LWF explained this behavior as a semantic consequence of the syntactic preference that *one* be anaphoric to the larger N’ string (*red bottle*). If the children had allowed *one* to be anaphoric to N^0 (bottle), they would have behaved similarly to the baseline condition and had a preference for the new bottle they hadn’t seen before. Since infants preferred the larger N’ string (as adults do) and this larger N’ string could not be classified as N^0, LWF concluded that the 18-month olds have the syntactic knowledge that *one* is anaphoric to N’ strings in general.

3.3.4. Sparse Data for Anaphoric *One*

In order to determine whether children’s knowledge could have been acquired on the basis of experience with the relevant forms and structures, LWF conducted a corpus analysis on child-directed speech. The important empirical question was how frequently data appeared in child-directed speech that signaled that *one* was anaphoric to N’ instead of N^0. If the data were not frequent, learning this syntactic knowledge would be difficult. The distribution LWF found is displayed in table 3.1 below.

Total Data in Corpus	Total # with anaphoric *one*
54,800	792
Data Type	# of data points
Unambiguous	2
“*Jack wants a red ball, but Lily doesn’t have another one for him*.” (Lily doesn’t have another ball with the property red.)	
Type I Ambiguous	36
“*Jack wants a red ball, and Lily has another one for him.*” (Lily has another red ball for Jack.)	
Type II Ambiguous	750
“*Jack wants a ball, and Lily has another one for him.”* (Lily has a ball with any number of properties.)	
Ungrammatical	4
“*…you must be need one.*” (Adam19.cha, line 940)	

Table 3.1. The distribution of utterances in the corpus examined by LWF.

All data are defined by a pairing of utterance and environment. I will now elaborate on the pairings for each data type. Unambiguous antecedent data have the following form:

(4) Unambiguous antecedent example
Utterance: “Jack wants a red ball, but Lily doesn’t have another one for him.”
Environment: Jack wants a red ball, but Lily doesn’t have another red ball – she has another ball with different properties.

Because Lily does indeed have a ball, the antecedent of *one* cannot be *ball*. However, Lily’s ball is not red, so the antecedent of *one* can be *red ball*. Since *red ball* can only be classified as N’, these data are unambiguous evidence that *one* can be anaphoric to N’.

An example of this type taken from the Adam corpus in CHILDES (MacWhinney, 2000) is given here. (Adam40.cha, line 890)

(5) CHI: Do you have another flat tire?
MOT: No. I don’t think I have one.

In this context, the mother had a tire, but not a flat tire, so the antecedent of *one* is unambiguously *flat tire*.

Type I ambiguous antecedent data have the following form:

(6a) Type I ambiguous antecedent example
Utterance: “Jack wants a red ball, and Lily has another one for him.”
Environment: Lily has a ball for Jack, and it has the property red.

(6b) Type I ambiguous antecedent example
Utterance: "Jack wants a red ball, but Lily doesn't have another one for him ."
Environment: Lily doesn't have another ball at all.

For data of the form in (6a), Lily has a ball, so the antecedent of *one* could be *ball.* However, Lily also has a ball that is red, so the antecedent of *one* could be *red ball.* Because *ball* could be classified as either N' or N^0, these data are ambiguous between *one* anaphoric to N' and *one* anaphoric to N^0.

An example of this type taken from the Adam corpus in CHILDES (MacWhinney (2000)) is given here (Adam01.cha, line 291).

(7) MOT: That's a big truck.
MOT: There goes another one.

In this context, *one* could be taken to refer to either *truck* or *big truck*.

For data of the form in (6b), Lily does not have a ball – but it is unclear whether the ball she does *not* have has the property red. For this reason, the antecedent of *one* is again ambiguous between *red ball* and *ball,* and *one* could be classified as either N' or N^0. There were no examples in either Adam or Nina's corpus of this form.

Type II ambiguous antecedent data have the following form:

(8a) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has various properties.

(8b) Type II ambiguous antecedent example
Utterance: "Jack wants a ball, but Lily doesn't have one for him."
Environment: Lily does not have another ball.

For both forms of type II ambiguous data, the antecedent of *one* must be *ball*. However, since *ball* can be classified as either N' or N^0, such data are ambiguous with respect to what *one* is anaphoric to.

An example of this type taken from the Adam corpus of CHILDES (MacWhinney (2000)) is given here (Adam01.cha, line 566).

(9) CHI: my pillow my
MOT: That's a good one to jump on.

Because there are no modifiers in the antecedent, *my pillow*, this data is uninformative about the structure of *one*.

There were no examples in either Adam or Nina's corpus of the form (8b).

Ungrammatical data involve a use of anaphoric *one* that is not in the adult grammar, such as in (10):

(10) Ungrammatical antecedent example
Utterance: "…you must be need one."

Since the utterance is already ungrammatical, it does not matter what environment it is paired with. The child will presumably be unable to resolve the reference of *one*. Such data is therefore noise in the input.

The vast majority of the anaphoric *one* input consists of type II ambiguous data (750 of 792, 94.7%). Type I ambiguous data makes up a much smaller portion (36 of 792, 4.5%). Ungrammatical data are quite rare (4 of 792, 0.5%), and unambiguous data rarer still (2 of 792, 0.25%). Since LWF considered unambiguous data as the only informative data, they concluded that such data seemed far too sparse to definitively signal to a learner that *one* is anaphoric to N'.

This seems in line with theory-neutral estimations of the quantity of data required for acquisition by a certain age (Legate & Yang, 2002). Specifically, other linguistic knowledge acquired by 20 months required at least 7% unambiguous signatures in the available data (Yang (2004) referencing Pierce (1992)). At least 1.2% unambiguous data was required for acquisition by 36 months (Yang (2004) referencing Valian (1991)). So, independent of what acquisition mechanism is assumed, having 0.25% unambiguous data makes it unlikely that the learner would be able to acquire the correct interpretation of anaphoric *one* by 18 months.

LWF's experimental results, however, suggested that 18-month olds know that *one* is anaphoric to N'. They therefore claimed that such knowledge does not need to be learned. Instead, the learner would have other innate biases that would allow this knowledge to be derived from the data available. One possibility (cf. Hornstein & Lightfoot (1981), Baker (1979)) would be that the child is constrained only to hypothesize phrasal antecedents for pronouns. Thus, once the child identified *one* as a pronominal form, the possibility that it was anaphoric to N^o would simply never be considered as a potential hypothesis.

3.4 Learning Anaphoric One

3.4.1 Suggestions for Learning that *One* is Anaphoric to N'

Two replies to LWF made suggestions for how this syntactic knowledge could be learned from the available data. The first reply by Akhtar et al. (2004) noted that even if the percentage of unambiguous data is quite small, 18-month olds have still been exposed to an estimated 1,000,000 utterances; this should yield a larger quantity of unambiguous data than the LWF corpus analysis obtained. So, a learner using only unambiguous data would encounter more unambiguous examples by 18 months. Still, the overall percentage of unambiguous data remains quite small (0.25%).

However, it is unlikely that this is even a fair estimate of the amount of data that the child has been exposed to. This is because much of the first year of life is spent learning phonological and lexical properties of the language which would be prerequisites to learning syntax. To derive a fairer estimate of the amount of relevant data an 18-month old might have been exposed to, I assume that learning the syntactic and semantic properties of *one* can only commence once the child has some repertoire of syntactic categories. Thus, I estimated that the learning period begins at 14 months because there is experimental data supporting infant recognition of the category Noun and the ability to distinguish it from other categories such as Adjective at this age (Booth & Waxman, 2003). If learners hear approximately 1,000,000 sentences from birth until 18 months, they should hear approximately 278,000 sentences of input between 14 months and 18 months. The adjusted expected distribution of anaphoric *one* data is displayed in table 3.2.

Total Data before 18 months	Total # with anaphoric *one*
~278,000	4017
Data Type	# of data points
Unambiguous	10
"*Jack wants a red ball, but Lily doesn't have another one for him.*" (Lily doesn't have another ball with the property red.)	
Type I Ambiguous	183
"*Jack wants a red ball, and Lily has another one for him.*" (Lily has another red ball for Jack.)	
Type II Ambiguous	3805
"*Jack wants a ball, and Lily has another one for him."* (Lily has a ball with any number of properties.)	
Ungrammatical	19
"*…you must be need one.*"	

Table 3.2. The expected distribution of utterances in the input to learners between 14 and 18 months.

Perhaps the most striking feature of this distribution is that there are still pitifully few unambiguous data points available. With only 10 chances to hear unambiguous data (on this estimate), a learner could well miss out due to fussiness, distraction, or other vagaries of toddler life. Moreover, this is still 0.25% of the anaphoric *one* data, which is well below the estimate of the amount of unambiguous data needed to acquire knowledge by 36 months (estimated at 1.2%, Yang (2004)), let alone by 18 months.

R&G offer a solution: make use of the type I ambiguous data as well, which gives 183 additional data points (on this estimate). Using a Bayesian learning model that implements the size principle of Tenenbaum & Griffiths (2001), R&G demonstrate how a learner could use both unambiguous and type I ambiguous data to converge on the correct representation. I review their learning model in the next section.

3.4.2 A Regier & Gahl Bayesian Learner

The power of R&G's model comes from using indirect evidence available in the type I ambiguous data. This is an attractive strategy, since there are nearly 20 times as many type I ambiguous data as there are unambiguous data (183 to 10). This raises the useable data for the learner up to 4.8% (193 of 4017), which seems more in line with the amount required for acquisition as early as 18 months (Yang (2004)). The indirect evidence itself is derived solely from the environment in which type I ambiguous data are uttered – specifically, by the learner examining the distribution of the referents of *one*. For example, suppose the learner hears type I ambiguous data such as the example in (6a) (repeated below as (11)):

(11) Type I Ambiguous
Utterance: "Jack wants a red ball, and Lily has another one for him."
Environment: Lily has a ball for Jack, and it has the property red.

Since the adult preference is to choose the larger N' as the antecedent, the antecedent of *one* will nearly always be *red ball* and the referent of the NP containing *one* will have the property red. The learner is able to observe the simultaneous presence of the larger N' as potential antecedent (*red ball*) and a referent in the world of *one* with the property mentioned in the N' (red). Note that this observation requires the learner to have a very abstract notion of what to generalize over. It is insufficient to generalize over a single property such as "red" or "behind his back"; instead, the learner must generalize over "property mentioned in the N' antecedent".

Now, the connection between the N' antecedent and a referent with the property mentioned in the N' will be true for some portion of the type I ambiguous data.[12] Crucially, for R&G's model, it is *never* true that the referent of *one* definitively lacks the property mentioned in the N' antecedent (i.e. the referent of *one* is definitively not red when the antecedent is *red ball*). A Bayesian learner using the size principle is very sensitive to this fact in the following way:

(12) Bayesian Learner Logic

(a) For type I ambiguous data, suppose that the referent of *one* could have any property, and not necessarily have the property mentioned in the larger N' antecedent. Suppose also that the set of potential referents for an utterance like (11) is represented in figure 13.

[12] This reasoning will not work for type I ambiguous data of the form in (2b): "Jack wants a red ball, but Lily doesn't have another one for him", where Lily does not have a ball. This is because the learner cannot tell whether or not the ball Lily doesn't have has the property red. These data are therefore not useful as indirect evidence. Such data did not occur in the Adam and Nina corpora from which my estimates are derived.

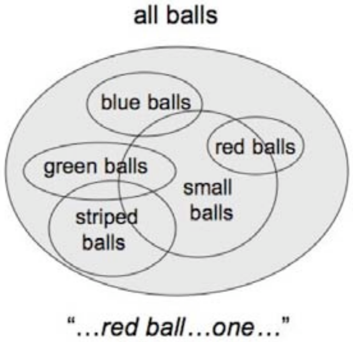


Figure 13. The set of potential referents for *one* in the world when an utterance such as "Jacks wants a red ball, and Lily has another one for him" is heard.

(b) The actual distribution of referents observed by the learner, however, is only a particular subset of all the possible referents.

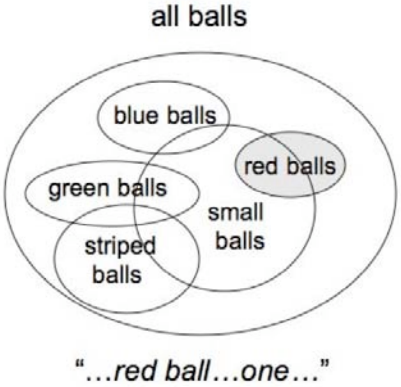


Figure 14. The observed set of referents for *one* when an utterance such as "Jack wants a red ball, and Lily has another one for him" is heard.

(c) It is highly unlikely that the referent of *one* is only ever a member of the subset if the referent could be any member of the superset. The Bayesian learner will therefore consider a restriction to the subset to be more and more probable as time goes on. This is the size principle of Tenenbaum & Griffiths (2001): if there is a choice between a subset and the superset, and only data from the subset is seen, the learner will be most confident that the subset is the correct hypothesis. Thus, the learner uses the lack of data for the superset as indirect evidence that the subset is correct.

The specific instantiation of the bias for the subset (red balls) given a single subset data point is based on the likelihood of encountering that subset data point. The likelihood of choosing a specific member of the subset (a red ball) is higher if members can be drawn only from the subset (red balls), as opposed to if members can be drawn from the superset (all balls). This occurs because the superset necessarily has more members to choose from, and therefore there is a lower probability of choosing a specific subset member.

The amount of bias a subset data point gives the subset depends on the relative sizes of the subset and superset. If the superset (all balls) has many more members than the subset (red balls), the likelihood of drawing a specific member from the subset (a red ball) when any member from the superset could have been chosen is low. The bias towards the subset (red balls) given a subset data point (a red ball) is then higher. In contrast, if the superset (all balls) has only a few more members than the subset (red balls), the likelihood of drawing a specific member from the subset (a red ball) when any member from the superset could have been chosen is higher. The bias towards the subset (red balls) given a subset data point (red ball) is then lower.

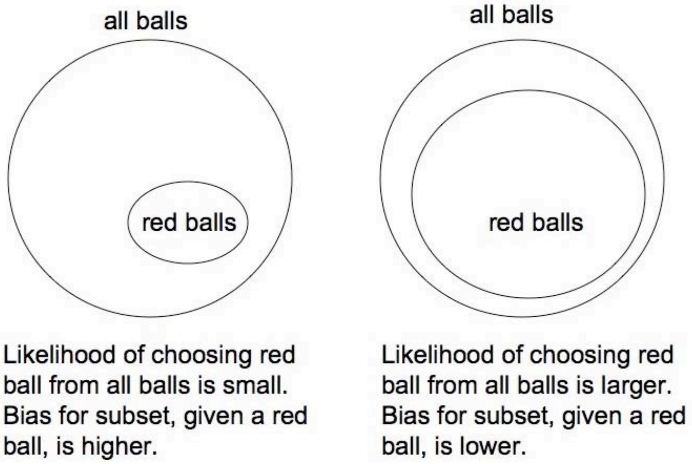

all balls
red balls
Likelihood of choosing red ball from all balls is small. Bias for subset, given a red ball, is higher.
all balls
red balls
Likelihood of choosing red ball from all balls is larger. Bias for subset, given a red ball, is lower.

Figure 15. Comparison of different ratios of superset to subset, the likelihood of choosing a member of the subset, and the effect on subset bias.

(d) Once the learner is biased to believe that there is a restriction to the subset of referents described by the property mentioned in the N' (*red* in *red ball*), the learner then assumes that the correct antecedent is, in fact, the larger N'.[13] Since the larger N' cannot be classified as N^0, the learner then knows that *one* always has an N' antecedent.

(e) For the LWF experiment, a Bayesian learner would have converged on the subset of red bottles as the potential referents of *one* in the test utterance. Given a choice between a red and a non-red bottle, the Bayesian learner therefore looks at the bottle that belongs to the correct subset: the red bottle.

A great strength of the R&G model is that the bias to choose the subset, given indirect evidence, does not need to be explicitly assumed. Instead, it falls out neatly from the mathematical implementation of the Bayesian learning procedure itself that uses the size principle of Tenenbaum & Griffiths (2001).

[13] R&G's model demonstrates how this could happen after very few type I ambiguous data.

However, as I noted before, the model implemented in the R&G study still harbors two implicit biases about domain-specific data filters on the learner's intake. The first bias is that only unambiguous and type I ambiguous data are used; type II ambiguous data are ignored even though they may also provide indirect evidence to a Bayesian learner. The second bias is that only semantic data (the referents of *one*) are used to converge on the syntactic knowledge of what *one* is anaphoric to; syntactic data are ignored.

In the remaining sections of the chapter, we will see that stripping away these two biases (and thus creating an unbiased learner truer to the spirit of R&G's proposal) leads to markedly different results from those of R&G. Specifically, once we remove these two biases, we will discover that a Bayesian learner will *not* learn that *one* is anaphoric to N' with high probability and will *not* choose the adult interpretation of the larger N' constituent with high probability when there is a choice between N' constituents. So, this unrestricted Bayesian learner will (a) have a preference for the wrong syntactic analysis (N^0) and (b) a preference for the wrong semantic interpretation (smaller N' (ball): do not require the referent to have the property mentioned in the antecedent), even if the correct syntactic analysis is chosen.

The benefit that comes from using indirect negative evidence to shift the majority of the probability to the subset in the hypothesis space is tempered by the link between the two levels of representation. Specifically, the semantic interpretation is a projection from the syntax. If indirect learning leads to the subset N^0 in the syntax, then the semantic preference to choose the interpretation consistent with the larger N' constituent when there is a choice between two N' constituents will not be helpful to the learner in most cases. This is simply because the learner will not choose the N' analysis very often, and so will have no need to access the semantic interpretation preference. Thus, the existence of multiple levels of representation reduces the efficacy of this kind of learner.

3.5 An Equal-Opportunity Bayesian Learner

3.5.1 Introducing the Equal-Opportunity Bayesian Learner

I will refer to the unrestricted domain-general learning model as the Equal-Opportunity (EO) Bayesian Learner since it removes the two implicit biases of R&G's Bayesian learner and so gives equal treatment to all data. First, it denies privileged status to a subset of the data and instead uses all the data available: unambiguous, type I ambiguous, and type II ambiguous. Second, it denies privileged status to semantic data – syntactic and semantic data are both used to shift probability between opposing hypotheses. There is an intuitive logic to using both types of data, since one should presumably use syntactic data (among other kinds of data) to converge on syntactic knowledge.[14] This syntactic knowledge has semantic

[14] Note that even if we believed the knowledge about *one* was stated purely in semantic terms, the data that any grammar predicts will include both syntactic data (i.e. what the linguistic antecedent for *one* is) and semantic data (what the referent of *one* is). So, excluding either kind of data is an arbitrary restriction on the learner that would need to be justified. For this reason, the hypothesis to include both

consequences, which are displayed in the LWF experiment. If a Bayesian learning procedure, unconstrained by domain-specific filters, is to be an effective domain-general learning solution, it should correctly acquire knowledge that spans domains such as syntax and semantics as well as knowledge contained completely within these domains.

3.5.2 The Hypothesis Space

The hypothesis space is defined for both the syntactic and semantic domains. The syntactic domain contains hypotheses about what strings can be antecedents for *one*. Each hypothesis covers a set of strings, and is classified by the syntactic category that can generate all the strings in the hypothesis. The semantic domain is a projection of the syntactic domain and contains hypotheses about the interpretation of *one* (specifically what referents in the world *one* can refer to). Each hypothesis covers a set of referents, and is classified by what properties the referents in that set must have. In both domains, there are two hypotheses to choose from. Each hypothesis makes predictions about the data that will be encountered and, consequently, the elements that will be analyzable under that hypothesis.

For each domain, the elements analyzable by one hypothesis are a subset of the elements analyzable by the other. For syntax, the hypotheses under consideration are (a) that *one* is anaphoric to strings that are classified as N^0 and (b) that *one* is anaphoric to strings that are classified as N'. Every string in N^0 can also be classified as N' but there are strings in N' that cannot be classified as N^0. Therefore, the strings that comprise the N^0 set are a subset of the strings that comprise the N' set.

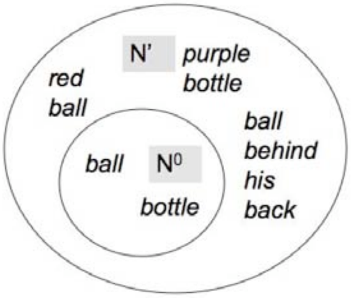


Figure 16. The syntax hypothesis space, N^0 vs. N'. All the elements in the sets are strings that are possible antecedents of *one*. Every string classified as N^0 can also be classified as N'. In addition, there are strings in N' that are not in N^0, and so the N^0 set is a subset of the N' set.

For the semantic interpretation, the referents of *one* could have the restriction that they must have the property named by the modifier; alternatively, the referents of *one* could have no restriction on what property they have. Since the modifier is linguistically not part of the N^0 (recall figure 11) and instead is part of the N' phrase,

syntactic and semantic data does not rely on a particular specification of knowledge about anaphoric *one*.

I will refer to the property named by the modifier as the N'-property. I will refer to referents with no restrictions as being any-property referents, since these referents can have any property (though of course they must still be instances of the noun in the antecedent, e.g. balls, if the antecedent is *red ball*). So, in the semantic domain, the two hypotheses under consideration are (a) that the referent of *one* is restricted to have the N'-property and (b) that the referent of *one* can have any property (is not restricted to have the N'-property).

Just as in the syntactic domain, the elements predicted by one hypothesis are a subset of the elements predicted by the other (see figure 17). Every referent that has the N'-property (say red for *red ball*) is a member of the N'-property set. By definition, every member of the N'-property set is also a member of the any-property set, since the N'-property is one of the properties available for objects to have. However, there are members of the any-property set (say green balls for the linguistic antecedent *red ball*) that do not have the N'-property (red). So, since all the members of the N'-property set are members of the any-property set, the N'-property set is a subset of the any-property set. Moreover, some members of the any-property set are *not* members of the N'-property set. So, the any-property set is a superset of the N'-property set in the semantic domain.

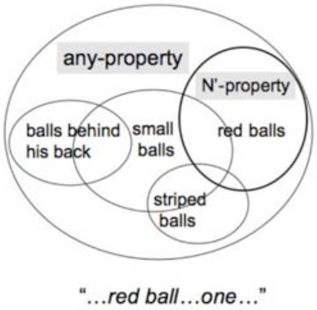


Figure 17. The semantic hypothesis space, N'-property vs. any-property. Any-property is a superset of N'-property. Note that in order to define the sets (N'-property vs. any-property), the utterance must be used to determine the salient property that the referent of the antecedent has. The salient property can be determined from the linguistic antecedent of *one*.

The difficulty for a Bayesian learner becomes apparent when we examine how the two prediction spaces defined by the hypotheses are connected. Specifically, in the syntactic domain, the relative complement of the subset in the superset (the set of strings that are in the superset but not the subset, such as *red ball*) is linked to the subset in the semantic domain; the subset in the syntactic domain is linked to the superset in the semantic domain. For ease of exposition, I will refer to the relative complement of the subset in the superset as the "exclusive superset".

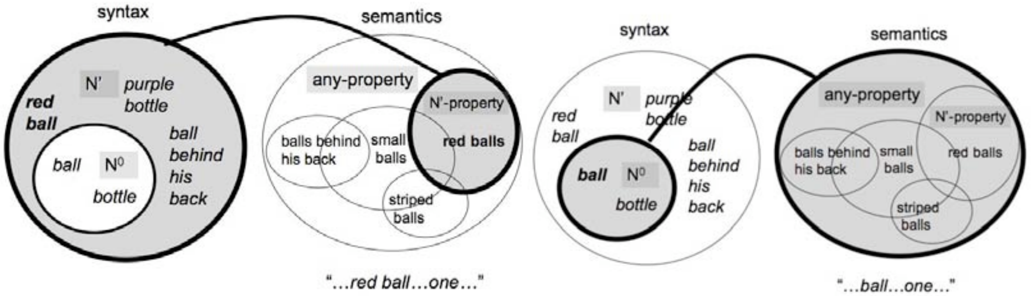


Figure 18. In the syntactic domain, the exclusive superset is linked to the subset in the semantic domain. The subset of the syntactic domain is linked to the superset in the semantic domain.

This is due to the compositional property of syntactic representations: larger syntactic constituents (such as the N' *red ball*) have meanings that are restrictions on the meanings (and so the referents) of their constituent subparts. In syntax, the strings in the exclusive superset (e.g. *red ball*) designate a subset of referents in the semantics (e.g. the red balls); the strings in the subset of the syntax (e.g. *ball*) designate the superset of referents in the semantics (e.g. all balls).

Because the syntactic and semantic representations are linked in this fashion, a Bayesian learner that relies on indirect evidence to shift probability towards the subset will receive conflicting information from across the two domains. For instance, the learner will encounter ambiguous data that favors the syntactic subset (the wrong answer for English anaphoric *one*). The learner will also encounter ambiguous data that favors the semantic subset which is linked to the exclusive superset in the syntax that implicates N' (the correct answer for English anaphoric *one*). However, this will not negate the aforementioned syntactic evidence that favors the syntactic subset N^0. Yet, the learner shouldn't ignore available syntactic information since anaphoric *one* has a representation at the syntactic level. Thus, we can see that an unrestricted Bayesian learner that uses all available data (syntactic and semantic) will need to overcome conflicting information across domains in order to converge on a high probability for the correct representations of anaphoric *one*.

It is important to recognize that the problem of linked hypothesis spaces extends far beyond the particular case of anaphoric *one*. Because syntactic structures are semantically compositional, this problem will persist across the acquisition of any aspect of the grammar that depends on the link between syntax and semantic reference.

3.5.3 EO Bayesian Learning

The EO Bayesian learning model uses Bayesian reasoning to update the learner's confidence in each of two alternative hypotheses. The implementation I will use differs from the R&G learner by being more conservative about updating the probabilities of the competing hypotheses. I will first describe the R&G Bayesian implementation, and then describe the implementation I will use here. I detail the learning process independently for each of the two domains (syntax and semantics) that are relevant for determining the appropriate structure of anaphoric *one*. I then describe how to implement the updating algorithm, given that these two domains are linked.

3.5.3.1 The R&G Bayesian Learner Implementation

The R&G learner is quite liberal about shifting probability to the superset hypothesis: a *single* piece of data for the exclusive superset is enough to shift *all* the probability to that hypothesis. However, as we have seen, the correct hypothesis for English anaphoric *one* is in the subset in the semantic domain: the learner should prefer the larger N' constituent, e.g. *red ball*, and thus restrict referents to those that have the N'-property, e.g. red balls. The success of this learner for converging on the correct semantic hypothesis for anaphoric *one* relies on the assumption that there will never be unambiguous data for the semantic superset.

Recall that the semantic superset hypothesis is that *one* refers to an object that does not need to have the property mentioned in the linguistic antecedent. This is the any-property hypothesis. Unambiguous data for the superset would be an utterance where *one* refers to an object that does *not* have the property mentioned in the antecedent. For instance, if the utterance is "…red ball…one…", unambiguous superset data would be the situation where the referent of *one* does not have the property 'red', e.g. it is a purple ball.

It is crucial for R&G's model that this type of data never occurs, though it is entirely possible that the learner might encounter this type of data as noise. If the referent of *one* in the above utterance was a purple ball (perhaps by accident), the new probability for the subset hypothesis (the N'-property hypothesis) in the semantic domain would be 0. I detail why this occurs below.

Suppose that we refer to the probability that the N'-property hypothesis is correct as $p_{N'\text{-}prop}$. Suppose the learner initially has no bias for either semantic hypothesis, and so the initial probability of $p_{N'\text{-}prop}$ is 0.5 before any data is encountered. This probability will increase as each piece of ambiguous (subset) data is observed, due to the size principle which biases the learner to favor the subset hypothesis if ambiguous data is observed.

Let *u* be a piece of unambiguous data for the superset hypothesis, where the utterance is "…red ball….one…" and the referent of *one* is a non-red ball. The learner now calculates the updated probability that the N'-property hypothesis is correct, using Bayes' rule. The updated $p_{N'\text{-}prop}$ given the observation of *u* is represented as the conditional probability p(N'-prop| *u*). To calculate this probability, we use Bayes' rule.

(13) Calculating the conditional probability p(N'-prop| *u*) using Bayes' rule

$$\text{p}(\text{N}'\text{-prop} \mid u) \;\propto\; \text{p}(u \mid \text{N}'\text{-prop}) * \text{p}(u)$$

The probability p(*u*|N'-prop) is the likelihood of observing the unambiguous superset data *u*, given that the N'-property hypothesis is true. In this case, the referent of *one* in *u* specifically doesn't have the N'-property ('red'). Therefore, it could not possibly be generated if the N'-property hypothesis was true, since the N'-property hypothesis requires the referent of *one* to have the property mentioned in the linguistic antecedent. So, the probability of observing *u* if the N'-property hypothesis is true (p(*u*|N'-prop)) is 0.

We substitute this value into the equation in (13) to get $\text{p(N'-prop} \mid u) \propto 0 * \text{p}(u) = 0$. Therefore, the updated probability for $p_{\text{N'-prop}}$ after seeing a single piece of unambiguous superset data *u* is 0, no matter what the previous probability of $p_{\text{N'-prop}}$ was.

Since this is not terribly robust behavior for a learner, I have adapted the Bayesian updating approach described by Manning & Schütze (1999) to generate a more conservative Bayesian updating approach, detailed in the previous chapter. Unlike the liberal R&G model, the learner using this more conservative approach shifts probability much more slowly between hypotheses. Only after observing a vast majority of evidence for one hypothesis would a conservative Bayesian learner shift the vast majority of the probability into that hypothesis.

3.5.3.2 Updating the Syntax Hypotheses

Recall that there are two hypotheses under consideration in the syntactic domain: the N' hypothesis and the N^0 hypothesis. The N' hypothesis takes the antecedent of *one* to be a constituent of the category N'; the N^o hypothesis takes the antecedent of *one* to be a constituent of the category N^0.

I represent the probability that the N' hypothesis is correct with $p_{N'}$. Because there are only two hypotheses in the hypothesis space, and because probabilities range from 0 to 1, the probability that the N^0 hypothesis is correct is $1 - p_{N'}$. I set the initial value of $p_{N'}$ before the learner has observed any data to 0.5 as an instantiation of the assumption that both hypotheses are equiprobable.

The update function requires a single parameter *t*, which represents the total amount of data expected during the learning period, as described in the previous chapter, and can be thought of as the total amount of change the real learner's brain is allowed to undergo before settling into the final state. In the simulated learner here, I quantify that amount of change as the total estimated amount of useable data available during the learning period (4017 data points, if using all available data). Of course, the value of *t* is essentially arbitrary, but in order to model this learning process, it needs to be estimated. The model uses *t* to determine how much probability shifting should be done, given a single piece of data. If *t* is small, only a small number of changes are allowed and each piece of data shifts the probability quite a lot; conversely, if *t* is large, a large number of changes are allowed and each piece of data shifts the probability a smaller amount. The value of *t* I use here will allow the modeled learner to converge as close as possible to an endpoint (e.g. $p_{N'} \approx 1.0$). In this way, I hope to estimate the best-case scenario for this kind of learner. While the *t* estimate presented here seems fair, I present a range of possible *t*-values in the results section. What we will see there is that the size of t does not influence the final probability of the correct interpretation of anaphoric *one*.

The exact update functions for $p_{N'}$ depend on the data type observed – unambiguous, type I ambiguous, or type II ambiguous. Unambiguous and type I ambiguous data cause the learner to use the function in (14a), which is essentially an implementation of the indirect negative evidence update function used by the R&G

model. Type II ambiguous data, which were not considered by the R&G learner, cause the EO Bayesian learner to use the function in (14b).

(14a) Update function for unambiguous and type I ambiguous data
Utterance: "…red ball…one…"
World: referent has the property red (unambiguous & some type I ambiguous) or it is unknown if referent has the property red (some of type I ambiguous)

$$\mathrm{p}_{N'} = \frac{\mathrm{p}_{N'\,\mathrm{old}} * t + 1}{t + 1}$$

(14b) Update function for type II ambiguous data
Utterance: "…ball…one…"
World: referent has various properties (type II ambiguous)

$$\mathrm{p}_{N'} = \frac{\mathrm{p}_{N'\,\mathrm{old}} * t + \mathrm{p}_{N'|\,a}}{t + 1}$$

The update function for unambiguous data is derived by using the mathematical framework laid out in the previous chapter. To briefly summarize, a binomial distribution centered at $\mathrm{p}_{N'}$ is used to approximate the learner's expectation of the distribution of the data to be observed. Data points from this distribution fall into two classes: they either have the "property" of being an N' data point or they do not have this property (and are instead N^0 data points). If $\mathrm{p}_{N'}$ is 0.5 (as it is initially), the learner expects half the informative data points to be N' data points. Using the derivations described in the previous chapter, we can then derive equation (14a) for updating $\mathrm{p}_{N'}$.

An intuitive interpretation of the unambiguous data update function is that the numerator represents the learner's confidence that the observed unambiguous N' data point *u* is a result of the N' hypothesis being correct; the denominator represents the total data observed so far. Thus, 1 is added to the numerator because the learner is fully confident that *u* indicates the N' hypothesis is correct; 1 is added to the denominator because a single data point has been observed.

Unambiguous data signal that the N' hypothesis is correct (in that only the N' hypothesis could have produced *u*) and so should be treated with full confidence by the learner. In contrast, the type I ambiguous data do not indicate that only the N' hypothesis could have produced *u* – these data are *ambiguous* between the N^0 and N' hypotheses. Thus, a smaller value should be added to the numerator for such data to indicate less than full confidence that only the N' hypothesis could have produced *u*.

However, I will allow the Bayesian learner to treat the type I ambiguous data with full confidence in the N' hypothesis. I make this allowance for two reasons. First, I know of no principled way to reasonably estimate how much confidence should be associated with a type I ambiguous data point. Second, this allowance is generous towards the Bayesian learner because it allows the model to overestimate the confidence the learner has in the N' hypothesis. If I was less generous and lessened the confidence in the type I ambiguous data, the probability of N' would only be lower than what I present here. As we will see below, even with this generous estimate, the learner will fail to assign sufficient probability to the N' hypothesis.

The update function for type II ambiguous data (14b), which comprise 3805 of the data points, depends on the prior probability that N' is the correct hypotheses ($p_{\text{N' old}}$), *t,* and a confidence value ($p_{\text{N'}\,|\,a}$). The intuitive interpretation for this function remains the same as the interpretation for the function in (14a): the numerator represents the learner's confidence that the observed ambiguous utterance-world pairing *a* is a result of the N' hypothesis being correct; the denominator represents the total data observed so far. Thus, a value less than 1 ($p_{\text{N'}\,|\,a}$) is added to the numerator because the learner is only partially confident that ambiguous data point *a* indicates the N' hypothesis is correct; and, 1 is added to the denominator because a single data point has been observed. The partial confidence value $p_{\text{N'}\,|\,a}$ depends on the likelihood that the utterance in *a*, which has only a noun string as the antecedent of *one* (ex: "…ball…one…"), would be produced if any N' string could have been chosen from the set of N' strings ($p_{\text{N from N'}}$).

The partial confidence value is the probability that *one* is anaphoric to N' in type II ambiguous data point *a*. This is equivalent to the probability that *one* is anaphoric to N' in general, given that *a* has been observed. I write it as Prob(N' | *a*) and calculate it by using Bayes' rule.

(15) $$\text{Prob}(\text{N'}|\,a) = \frac{\text{Prob}(a\,|\,\text{N'}) * \text{Prob}(\text{N'})}{\text{Prob}(a)}$$

I now describe the individual pieces of the right hand side of the equation in (15). Prob(*a* | N') is the probability of observing a type II ambiguous data point *a*, given that the N' hypothesis is true. Recall that a type II ambiguous data point has an utterance with a noun-only antecedent, such as "…ball…one…". The N' hypothesis states that the linguistic antecedent of *one* must be an N' constituent.

It is possible for a noun-only string to be an N' constituent: this is the situation where a noun-only string is chosen from the set of N' constituents, which consists of both noun-only strings ("ball", "bottle", etc.) and other strings that include modifiers ("red ball", "bottle in the corner", etc.). The probability we want is the probability of choosing a noun-only linguistic antecedent for *one* (such as in type II ambiguous utterance *a*), given the entire set of N' constituents. Suppose there are *n* noun-only strings and *o* other strings in the N' constituent set. I refer to the probability of choosing a noun-only string (such as "ball") as $p_{\text{N from N'}}$, and it is calculated below in (16).

(16) $$\text{Prob}(a\,|\,\text{N'}) = \frac{n}{n+o} = p_{\text{N from N'}}$$

Prob(N') is the current probability that the N' hypothesis is correct. This is simply $p_{\text{N'}}$.

Prob(*a*) is the probability of observing a type II ambiguous utterance *a*, no matter which hypothesis is correct. To calculate this value, we can sum the conditional probabilities of observing *a* for each hypothesis (Prob(*a* | N') + Prob(*a* |

N^0)) . If N’ is the correct hypothesis, the probability of observing *a* is Prob(*a* | N’) from above. If N^0 is the correct hypothesis, then the linguistic antecedent of *one* is an N^0 constituent, which is always a noun. In that case, the probability of observing a noun-only linguistic antecedent (such as in *a*) is 1. We can calculate Prob(*a*) in (17).

$$\text{(17) Prob}(a) \quad = \sum_{hypotheses} p_{hypothesis} * p(a \mid p_{hypothesis})$$
$$= p_{N'} * p(a \mid p_{N'}) + p_{N0} * p(a \mid p_{N0})$$
$$= p_{N'} * \frac{n}{n+o} + (1 - p_{N'}) * 1$$

Substituting these pieces back into the right hand side of the equation in (15), we obtain (18).

$$\text{(18) Prob}(N' \mid a) = \frac{(\frac{n}{n+o}) * p_{N'}}{p_{N'} * (\frac{n}{n+o}) + (1 - p_{N'}) * 1} = \frac{p_{N \text{ from } N'} * p_{N'}}{p_{N'} * p_{N \text{ from } N'} + (1 - p_{N'}) * 1} = p_{N' \mid a}$$

As we can see, the partial confidence value $p_{N' \mid a}$ depends only on $p_{N \text{ from } N'}$ and the current $p_{N'}$. This partial confidence value, which will be less than 1, is added to the numerator of the type II ambiguous data update function instead of 1. The larger $p_{N \text{ from } N'}$ is, the less biased the learner’s confidence is towards the subset N^0 hypothesis when a type II ambiguous data point is observed. This is because a higher $p_{N \text{ from } N'}$ signals that the superset N’ is not much larger than the subset N^0. So, the learner is not heavily biased towards the subset because the likelihood of choosing data point *a* from the subset is not much higher than the likelihood of choosing data point *a* from the superset. Thus, the more likely it is that a noun-only string could be chosen from the N’ constituent set, the less the N’ hypothesis is penalized when this type of data is seen.

The likelihood value $p_{N \text{ from } N'}$ is what allows the learner to retrieve information from the type II ambiguous data. The more unbalanced the ratio of noun-only strings to other strings in the N’ set, the stronger the effect of the size principle will be that biases the learner towards the subset N^0 hypothesis. Example (19) displays how much biasing occurs after a single piece of type II ambiguous data, assuming a current $p_{N'}$ of 0.5, a ratio of noun-only strings to total N’ strings of 0.25, and a *t* of 4017.

(19) Updated $p_{N'}$ after a single type II ambiguous data point *a*
Let $p_{N'} = 0.5$, $p_{N \text{ from } N'} = 0.25$, and $t = 4017$.
Updated $p_{N'} = .499925$ (a very slight bias for the N^0 hypothesis)

While the amount of bias towards the N^0 hypothesis is quite small, keep in mind that the majority of the data is type II ambiguous and so these small biases will add up over time.

3.5.3.3 Updating the Semantics Hypotheses

Recall that there are two hypotheses under consideration in the semantic interpretation domain that are projections from the syntactic domain: the N'-property hypothesis and the any-property hypothesis. The N'-property hypothesis requires the referent of *one* to have the property mentioned in the N' antecedent (e.g. red if the potential antecedent was *red ball*); the any-property hypothesis allows the referent of *one* to have any property. In this case, it's the N'-property hypothesis that represents the subset hypothesis. Thus, as above, the size principle will favor this hypothesis for any data that is compatible with both hypotheses.

I represent the probability that the N'-property hypothesis is correct with $p_{N'\text{-prop}}$. Because there are again only two hypotheses in the hypothesis space, the probability that the any-property hypothesis is correct is 1- $p_{N'\text{-prop}}$. I set the initial value of $p_{N'\text{-prop}}$ before the learner has observed any data to 0.5 as an instantiation of the assumption that both hypotheses are equiprobable.

The update function requires two parameters: *t* and *c*. As before, *t* represents the total amount of data expected during the learning period and is instantiated in this model as 4017, the estimated amount of data available during the learning period. The parameter *c* represents the number of properties (or *categories* of referents) in the world that the learner is aware of (e.g. red, striped, behind his back, etc.).

For the semantic domain, the data are divided according to how the properties of the referent of *one* compare to the salient property in the N' antecedent. The data types, representing the utterance-world pairings, are same-property, different-property, and unknown-property.

Same-property examples are those in which the potential antecedent of one mentions some property and the referent of one also has that property. Some of the data analyzed as type I ambiguous in the syntactic domain are same-property data. There are 183 or less data points of this form (because some portion of type I ambiguous are unknown-property data points).

(20a) Example of same-property data (syntax: type I ambiguous)

Utterance: "Jack wants a red ball, and Lily has another one for him."
World: Lily has another red ball for Jack.

The referent of *one* (the ball that Lily has) has the same property mentioned in the N' antecedent (red).

The data analyzed as unambiguous in the syntactic domain are also same-property data in the semantic domain. There are 10 data points of this form. Because these data necessarily include negation, seeing why they are same-property data is a bit complicated. Consider the example in (20b).

(20b) Example of same-property data (syntax: unambiguous)

Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."
World: Lily has a non-red ball for Jack.

The speaker in this situation is asserting the absence of a red ball. The referent of *one* is a red ball that is not present in the situation. Thus, the meaning of *one* includes the property mentioned in the antecedent.

Because the N'-property hypothesis depends on matching the property overtly mentioned in the modifier (e.g. *red* of *red ball*), type II ambiguous data are not informative for choosing between the two hypotheses. This is simply because there is no overtly mentioned modifier, as shown in (20c). Therefore, the semantic interpretation projection from the syntactic hypothesis space is a single hypothesis (the any-property hypothesis). Since the semantic domain only has one hypothesis for type II ambiguous data, no information can be inferred about the correct hypothesis when there is more than one semantic interpretation to choose. The learner therefore ignores the semantic hypothesis space when encountering type II ambiguous data.

(20c) Example of same-property data (syntax: type II ambiguous)

Utterance: "Jack wants a ball, and Lily has another one for him."

World: Lily has a ball with some property for Jack.

A different-property example is given in (21), when the potential antecedent has a property mentioned in the modifier (e.g. red of *red ball*), but the referent of *one* does not have this property. This situation would occur in rare cases, perhaps as noise or perhaps because of a pragmatic bias.

(21) Example of different-property data (syntax: type I ambiguous)

Utterance: "Jack likes a red ball, and Lily likes that one."

World: Lily likes a ball that is not red. (i.e., the referent of *one* is a non-red ball, even though the potential antecedent mentions the property *red*).

In this case, the semantic interpretation hypothesis unambiguously favored is the any-property hypothesis, since the data point is specifically in the exclusive superset of balls that do not have the N'-property (red). So, this kind of data strongly biases the learner towards the any-property hypothesis, the superset hypothesis in the semantic domain. That, in turn, biases the learner towards the subset in the syntactic domain (the smaller N' constituent, if the N' analysis is chosen). However, I will be generous and assume that this data does not occur in the EO Bayesian learner's dataset. This assumption will cause the EO Bayesian learner to overestimate the probability assigned to the N'-property hypothesis, $p_{N'\text{-prop}}$.

Finally, we come to the unknown-property data, as in (22).

(22) Example of unknown-property data (syntax: type I ambiguous)

Utterance: "Jack wants a red ball, but Lily doesn't have another one for him."

World: Lily has no ball for Jack.

In the examples in (22), the speaker is asserting the absence of a ball. The referent of *one* is a ball, with some unknown properties, that is not present in the situation. Thus, the referent of *one* may or may not include the property (red) mentioned in the potential antecedent.

A portion of type I ambiguous data consists of unknown-property data. Such data cannot be used for updating the probabilities of the opposing semantic hypotheses. However, I will be generous and allow R&G's assumption to hold true: none of the type I ambiguous data are of this form. Therefore, I will allow all type I ambiguous data to be of the form in (20a), which is an example of same-property data. This gives an overestimation of $p_{N'\text{-prop}}$, which is the subset in the semantic hypothesis space. Consequently, this will bias the learner towards the superset in the syntactic hypothesis space, N'. Thus, the model here will again overestimate the amount of probability the learner will assign to the correct hypothesis for the structure and interpretation of anaphoric *one*, given an utterance with more than one potential antecedent.

Table 3.3 represents the expected distribution of data for updating the semantic hypotheses in this model.

Total Data before 18 months	Total # with anaphoric *one*
~278,000	4017
Data Type	# of data points
Same-Property	10 + 183
"*Jack wants a red ball, and Lily has another one for him.*" (Lily has a red ball for Jack.) "*Jack wants a red ball, but Lily doesn't have another one for him.*" (Lily has a non-red ball for Jack.)	
Different Property	0
"*Jack likes this red ball, and Lily likes that one.*" (Lily likes a ball without the salient property that the antecedent referent has.)	
Unknown Property	0
"*Jack wants a red ball, but Lily doesn't have another one for him."* (Lily has no ball for Jack.)	

Table 3.3. The expected distribution of utterances in the input to the Bayesian learner for updating the semantics hypotheses. Note that the type II ambiguous data points are uninformative in the semantic interpretation domain, so those 3805 data points are ignored.

The exact update functions for $p_{N'\text{-prop}}$ depend on the data type observed. However, the only update function relevant for this model is the same-property update function (23), which is similar to its syntactic counterpart in (14b). In both cases, the subset hypothesis is favored upon encountering an ambiguous data point.

(23) Update function for same-property data

$$p_{N'\text{-prop}} = \frac{p_{N'\text{-prop - old}} * t + p_{N'\text{-prop} \mid s}}{t + 1}$$

The same-property update function is derived using the same reasoning as the type II ambiguous update function in the syntactic domain. We again have two hypotheses (N’-property and any-property), and so we can use a binomial distribution to approximate the learner’s expectation of the distribution of data to be encountered. The binomial distribution is centered at $p_{N'\text{-prop}}$, so the learner’s expectation is about how many N’-property data points should be observed. To update $p_{N'}$ after seeing a single same-property data point *s*, we again follow the framework laid out in the previous chapter and calculate the maximum of the a posteriori (MAP) probability.

Like the type II ambiguous data update function in the syntactic domain, however, we will add a value smaller than 1 to the numerator. Intuitively, this smaller value represents the learner’s smaller confidence that the same-property data point *s* indicates that the N’-property hypothesis is correct. I call this smaller value the partial confidence value, and represent it as $p_{N'\text{-prop} \mid s}$.

The partial confidence value $p_{N'\text{-prop} \mid s}$ is the probability that the referent of *one* must have the N’-property mentioned in *s*. This is equivalent to the probability that the referent of *one* must have the N’-property in general, given that *s* has been observed. I write it as Prob(N’-prop | *s*) and calculate it by using Bayes’ rule.

(24) $$\text{Prob(N'-prop}|\, s) = \frac{\text{Prob}(s\,|\,\text{N'-prop}) * \text{Prob(N'-prop)}}{\text{Prob}(s)}$$

I now describe the individual pieces of the right hand side of the equation in (24). Prob(*s* | N’-prop) is the probability of observing a same-property data point *s*, given that the N’-property hypothesis is true. Recall that in a same-property data point, the referent of the antecedent of *one* must have the same mentioned property that the referent of *one* has. The N’-property hypothesis states that the referent of the antecedent of *one* must have the property described by the linguistic antecedent of *one*. Therefore, if the N’-property hypothesis is true, the probability of observing a same-property data point is 1.

(25) $\text{Prob}(s|\text{N'-prop}) = 1$

Prob(N’-prop) is the current probability that the N’-property hypothesis is correct. This is simply $p_{N'\text{-prop}}$.

Prob(*s*) is the probability of observing a same-property utterance *s*, no matter which hypothesis is correct. To calculate this value, we sum the conditional probabilities of observing *s* for each hypothesis (Prob(*s* | N’- prop) + Prob(*s* | any-prop)) . If N’-property is the correct hypothesis, the probability of observing *s* is Prob(*s* | N’-prop) from above. If any-property is the correct hypothesis, then there is no restriction on what property the referent of the linguistic antecedent of *one* has. I estimate the probability of that referent having the same property by chance as the referent of *one* as simply 1/*c*, where there are *c* properties in the world. I calculate Prob(*s*) in (26).

$$\text{(26) Prob}(s) \quad = \quad \sum_{hypotheses} p_{hypothesis} * p(s \mid p_{hypothesis})$$

$$= \text{p}_{\text{N'-prop}} * \text{p}(s \mid \text{p}_{\text{N'-prop}}) + \text{p}_{\text{any-prop}} * \text{p}(s \mid \text{p}_{\text{any-prop}})$$

$$= \text{p}_{\text{N'-prop}} * 1 + (1 - \text{p}_{\text{N'-prop}}) * \frac{1}{c}$$

Substituting these pieces back into the right hand side of the equation in (24), we obtain (27).

$$\text{27) Prob(N'-prop}\mid s) = \frac{1 * \text{p}_{\text{N'-prop}}}{\text{p}_{\text{N'-prop}} * 1 + (1 - \text{p}_{\text{N'-prop}}) * \frac{1}{c}} = \frac{\text{p}_{\text{N'-prop}}}{\text{p}_{\text{N'-prop}} * + \frac{(1 - \text{p}_{\text{N'-prop}})}{c}} = \text{p}_{\text{N'-prop} \mid s}$$

As we can see, the partial confidence value $\text{p}_{\text{N'-prop} \mid s}$ depends only on *c* and $\text{p}_{\text{N'-prop}}$. This partial confidence value, which will be less than 1, is added to the numerator of the same-property data update function instead of 1. The larger *c* is, the larger the ratio between the any-property superset and the N'-property subset. The larger that ratio is, the more the learner is biased towards the subset hypothesis when encountering a same-property data point. Thus, when *c* is large, the learner's confidence in the N'-property hypothesis is high when encountering a same-property data point. So, the more properties there are in the learner's world, the more the N'-property hypothesis is rewarded when this type of data is seen. As for the denominator of the update function, we add 1 because a single data point has been observed.

3.5.4 The Updating Algorithm for Linked Domains

Recall that there is an inherent connection between the syntax and the semantic interpretation. In particular, the subset hypothesis in the syntax (N^0, or the smaller N' constituent) corresponds to the superset hypothesis in the semantics (any-property), and the exclusive subset in the syntax (larger N' constituents) corresponds to the subset (N'-property) in the semantics (figure 18). Given this arrangement of hypothesis spaces, any piece of data impacting a hypothesis in one domain should impact the corresponding hypothesis in the other domain by the same amount. I now provide a description of how I model this process.

First, suppose the learner receives an unambiguous or type I ambiguous data point (which have two strings as potential antecedents, e.g. *red ball* or *ball*). This data point can be analyzed in either domain, syntax or semantics. So, the learner chooses which one to analyze it in first. Then, the update functions described above are employed to determine the amount the probability that should be shifted within that domain. Next, the probability is shifted in the other domain by the same amount. See figure 19, which shows the learner analyzing the data in syntax and updating both syntax and semantics. Now, the learner analyzes the data point in the other domain, applies the update functions described previously to determine the amount the probability that should be shifted within this domain. Next, the probability is shifted in the other domain by the same amount. See figure 20, which shows the learner analyzing the data in the semantics and updating both semantics and syntax.

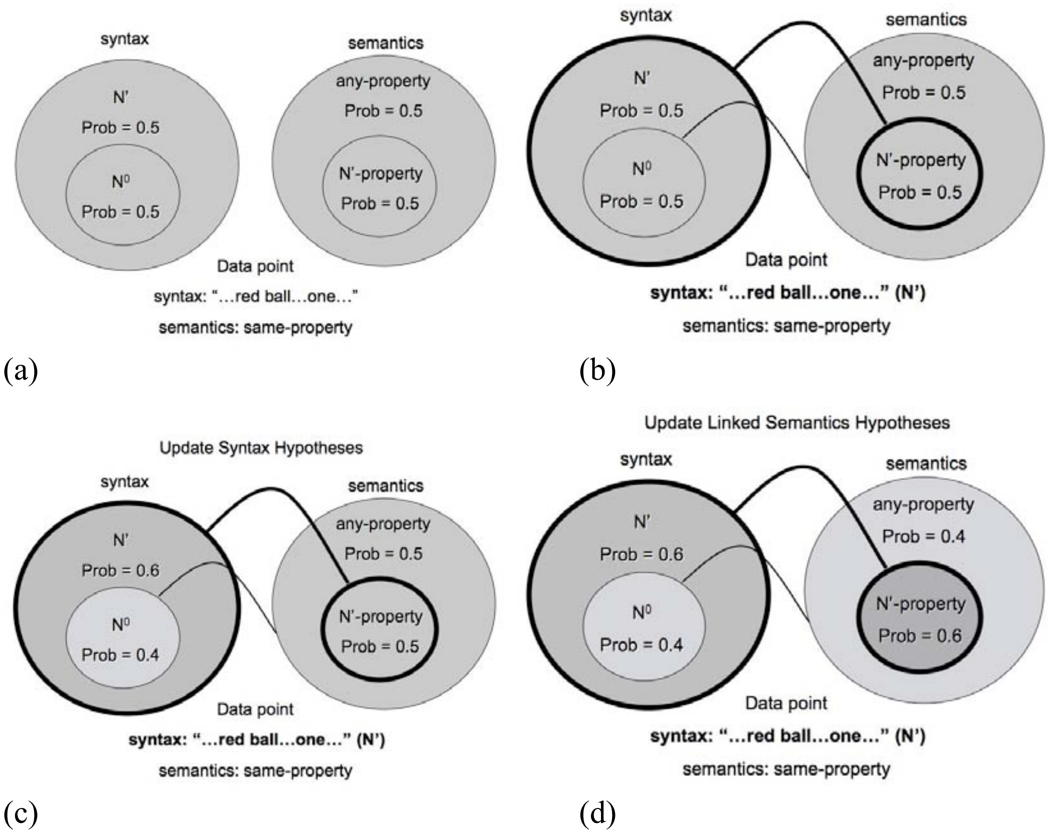
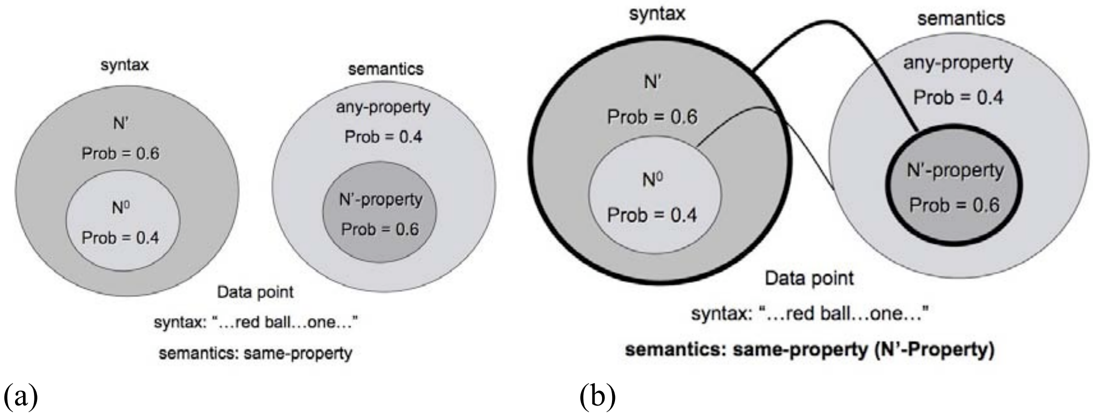


Figure 19. The learner encounters an unambiguous data point (a) and analyzes it first in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c) and the probability of the linked semantics hypotheses (d).



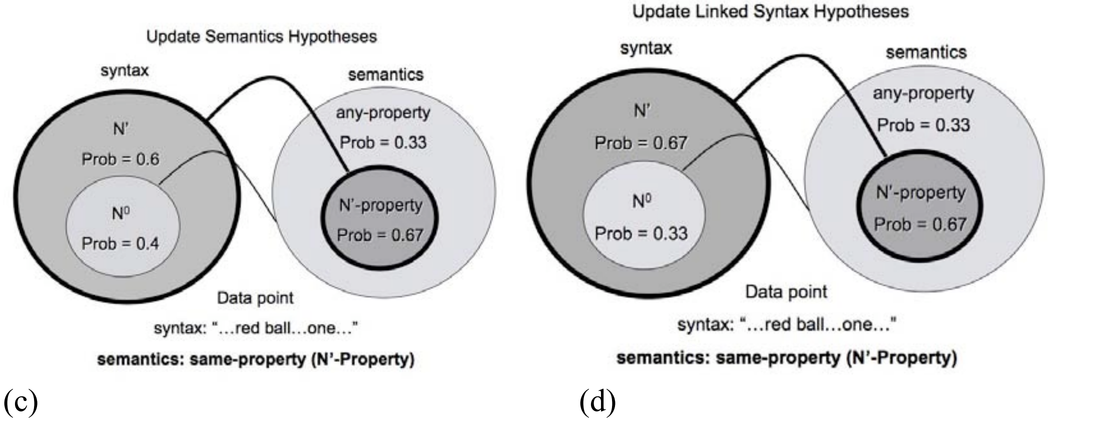


Figure 20. After analyzing the data point in the syntax domain and updating the probabilities across the domains, the learner then starts at the state in (a) and analyzes the data point in the semantics domain (b). Then, the learner updates the probability of the semantics hypotheses (c) and the probability of the linked syntax hypotheses (d).

The update process differs for a type II ambiguous data point, however. This is because there is only one string that is the potential antecedent (e.g. *ball*), and the projection from the syntax to the semantics leaves only one interpretation (any-property). Type II ambiguous data points are thus uninformative for the semantic interpretation domain. So, the learner simply updates in the syntax domain alone, as shown in figure 21. The semantic interpretation domain is ignored for this type of data.

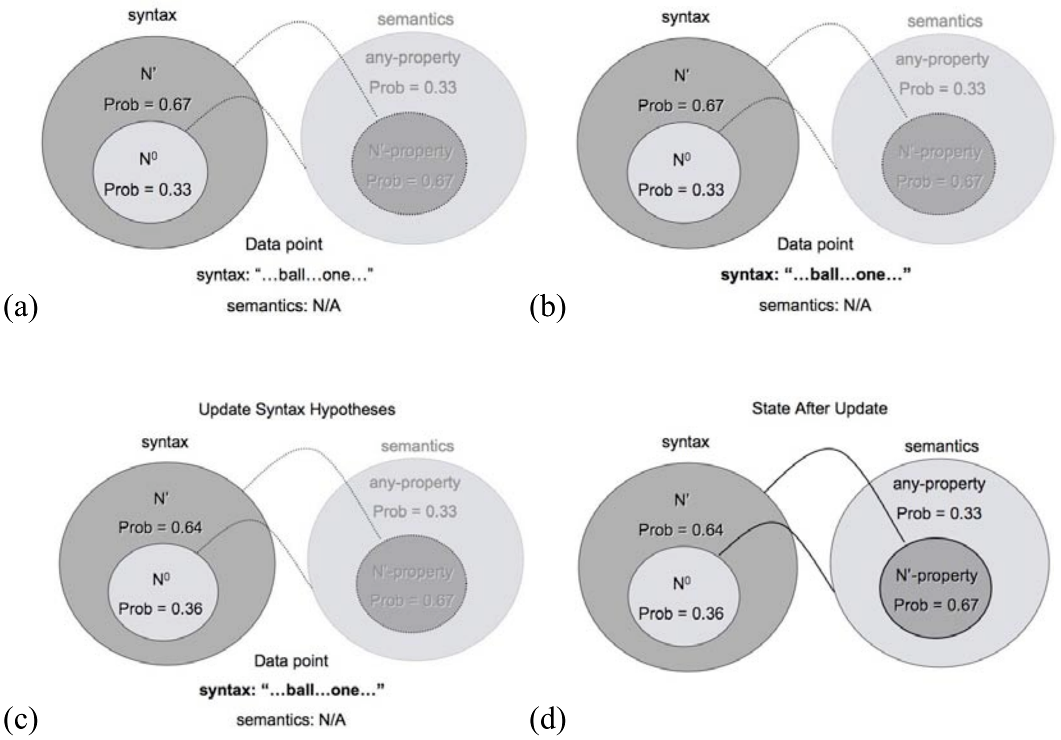


Figure 21. The learner encounters a type II ambiguous data point (a) and analyzes it in the syntactic domain (b), and then updates the probability of the syntax hypotheses (c). The final state after update is show in (d). Importantly, the semantic domain is not influenced by the type II ambiguous data point because there is only one semantic interpretation available for an antecedent with no modifiers (e.g. *ball*), the any-property hypothesis. The semantic domain is only influenced when there is more than one potential antecedent, leading to more than one semantic interpretation.

3.5.5 What Good Learning Would Look Like

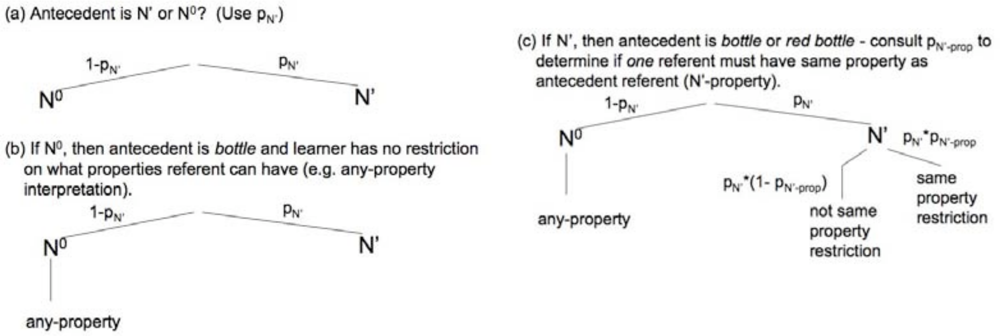
In the model, the learner initially assigns equal probability to the two hypotheses in each of the two domains: in the syntax, N^0 and N’, and in the semantics, N’-property (corresponding to the larger N’ constituent interpretation, e.g. *red ball*) and any-property (corresponding to the smaller N’ constituent interpretation, e.g. *ball*). The probability of choosing the preferred adult interpretation, given an utterance with two potential antecedents, depends on choosing the correct hypothesis in each domain. So, if the learner hears, “Look! A red bottle! Do you see another one?” (as in the LWF experiment), the interpretation of *one* is calculated as in (28), which is schematized in the decision tree in figure 22.

(28) Interpreting *one* in “Look! A red bottle! Do you see another one?”

(a) Determine if the antecedent of *one* should be N^0 or N’, using $p_{N'}$.

(b) If the antecedent is N^0, then the referent can have any-property.

(c) If the antecedent is N’, use $p_{N'\text{-prop}}$ to determine if the smaller N’ constituent interpretation (any-property) or the larger N’ constituent interpretation (N’-property) should be used.


(a) Antecedent is N' or N⁰? (Use p_N')
1-p_N'
p_N'
N⁰
N'
(b) If N⁰, then antecedent is bottle and learner has no restriction on what properties referent can have (e.g. any-property interpretation).
1-p_N'
p_N'
N⁰
N'
any-property
(c) If N', then antecedent is bottle or red bottle - consult p_N'-prop to determine if one referent must have same property as antecedent referent (N'-property).
1-p_N'
p_N'
N⁰
N'
p_N'*p_N'-prop
p_N'*(1- p_N'-prop)
any-property
not same property restriction
same property restriction

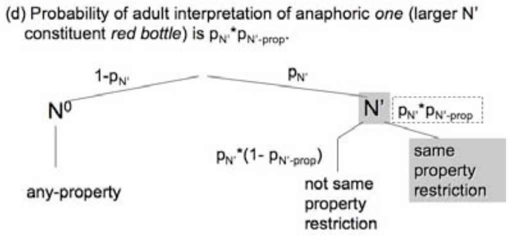

(d) Probability of adult interpretation of anaphoric *one* (larger N' constituent *red bottle*) is $p_{N'}*p_{N'\text{-prop}}$.
$1-p_{N'}$
$p_{N'}$
N^0
N'
$p_{N'}*p_{N'\text{-prop}}$
$p_{N'}*(1- p_{N'\text{-prop}})$
any-property
not same property restriction
same property restriction

Figure 22. Decision tree to interpret anaphoric *one* in utterances with more than one potential antecedent, such as “Look! A red bottle! Do you see another one?” The probability of having the adult interpretation (*one* = *red bottle*) is $p_{N'}*p_{N'\text{-prop}}$.

The probability of choosing the preferred adult interpretation (the larger N’ constituent is the antecedent of *one*) is the product of the probability of choosing the correct hypothesis in the syntax (N’) and that of choosing the correct hypothesis in the semantic interpretation (N’-property = larger N’ constituent): 0.500 * 0.500 = 0.250. Given that the end state should be a probability near 1, a good learning algorithm should have a trajectory like that illustrated in figure 23. In short, the learner should steadily increase the probability of choosing the preferred adult interpretation.

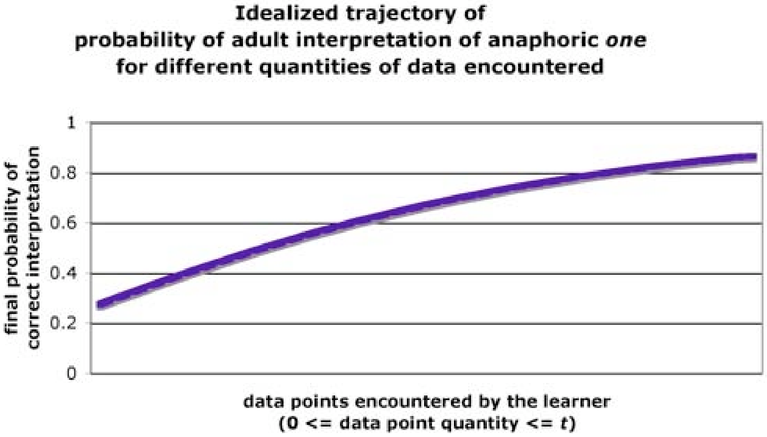


Figure 23. The idealized trajectory of the probability of the correct interpretation for anaphoric *one* as a function of the data points encountered by the learner.

3.5.6 Simulating an EO Bayesian Learner

Now that we have established how an EO Bayesian Learner learns and what the ideal learning outcome would be, we can simulate learning over our estimate of the set of data that 18-month olds have been exposed to. Each data point is analyzed in both the syntax and semantics domains, as relevant to the data type; and, each data point is classified for both syntax (unambiguous, type I ambiguous, or type II ambiguous) and semantics (same-property only, by generous assumption).

3.5.6.1 Syntax

The probability $p_{N'}$ is updated as each data point is observed. The model requires a value for $p_{N \text{ from } N'}$, the probability of choosing a noun-only string from the N' string set. This requires that we determine how many strings are in the N' set. There are two ways of doing this. First, we could allow a string to consist of individual vocabulary items ("bottle", "ball", "ball behind his back", etc.). Alternatively, we could allow a string to consist of individual categories (Noun, Noun PrepositionalPhrase, etc.). Recall that as $p_{N \text{ from } N'}$ increases, the ratio between superset size and subset size decreases and the N'-hypothesis is not penalized as much by a type II ambiguous data point. This means that a higher $p_{N \text{ from } N'}$ will generate a higher estimate for $p_{N'}$. Therefore, to be generous and maximize the model's estimate of $p_{N'}$, I choose the option that maximizes the value of $p_{N \text{ from } N'}$ and allow the strings in the N' set to consist of individual categories instead of vocabulary items. The number of categories is necessarily smaller than the number of vocabulary items in those categories, and so this yields a larger value for $p_{N \text{ from } N'}$.

Let the set of strings in N' = {Noun, Adjective Noun, Noun PrepositionalPhrase, Adjective Noun PrepositionalPhrase}.[15] The probability of producing a Noun string from this N' string set is 1/4 or 0.25. We can now look at the semantic domain.

3.5.6.2 Semantics

The probability $p_{N'\text{-prop}}$ is updated as each data point is observed. The model requires a value for *c*, the number of properties in the learner's world. Recall that as *c* increases, the ratio between the superset (any-property) and subset (N'-property) increases; the higher this ratio, the more the subset hypothesis (N'-property) is rewarded whenever a same-property data point is encountered. Data from the MacArthur CDI (Dale & Fenson, 1996) suggest that 14-16 months olds know at least 49 adjectives. Therefore, I estimate that an 18-month old learner should be aware of at least 49 properties in the world.[16]

Note however that it is unlikely all 49 properties to choose from would be represented in a given situation (nice balls vs. red balls vs. blue balls vs. pretty balls, etc.). Instead, a subset of the available categories the learner knows would be available in each case (perhaps as few as two: a red ball vs. a blue ball, for instance). So, assuming the learner considers the potential 49 properties the semantic referent in a given situation *could* have had will be an overestimation of the categories the learner actually considers. Because of this, the simulated learner will receive more bias towards the semantic subset (the correct interpretation of anaphoric *one*) than a real learner would. This will again yield an overestimation of a real learner's

[15] This is still a conservative estimate – there are likely to be additional category strings in N', such as Adjective Adjective Noun, because language is recursive. Additional strings would again lower $p_{N \text{ from } N'}$.

[16] In reality, there are still more properties due to the combination of adjectives (nice red, big striped) and prepositional phrases (nice…behind his back, big striped…in the corner). I will not consider the consequences of recursive modification here.

probability of choosing the more restricted referent set in the semantics, and thus an overestimation of the probability of the learner choosing the correct interpretation.

3.5.6.3 Linked Domain Updating

Recall that the update algorithm analyzes each data point in two domains and shifts the probability between the opposing hypotheses within each domain and across domains accordingly, as relevant. As we can see in figure 24, the learning trajectory as a function of the amount of data seen does not match our ideal learning outcome. In fact, as the learner encounters more data, the probability of the adult interpretation steadily drops to a final value of 0.171. This final value represents the product of the probability of the correct syntactic hypothesis ($p_{N'}$), which is 0.310 (1000 simulations, sd = .00377) and that of the correct semantic interpretation hypothesis ($p_{N'\text{-prop}}$), which is 0.551 (1000 simulations, sd = .00382).[17] Thus, based on the data observed, the learner is extremely unlikely to access the preferred adult interpretation for *one* (i.e., that *one* is anaphoric to strings described by N', and that the referent of *one* must have the N' property) in an utterance with two potential antecedents.

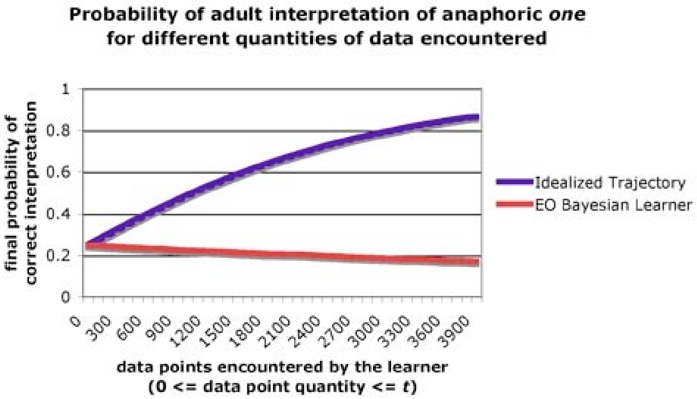


Figure 24. The EO Bayesian Learner's trajectory as a function of the amount of data encountered compared against the idealized trajectory for a learner.

3.5.6.4 Changing *t*

Recall that this model contains a parameter, *t*, which represents the amount of change the learner can undergo in the course of learning. I quantify this parameter as the number of data points the learner can use to update its probabilities. In my simulation, this was 4017, the number of data estimated during the learning period for an 18-month old. However, one might be concerned that the value of *t* might play a critical role in determining the final probability of converging on the correct

[17] Note that this value is obtained using the procedure in which the learner chooses at random whether to analyze the data point in the syntax first or in the semantics first for unambiguous and type I data. The same value is obtained if the learner always analyzes the data point in the syntax first and if the learner always analyzes the data point in the semantics first.

interpretation of anaphoric *one*. In figure 25, I show the final probability of converging on the adult interpretation of anaphoric *one* as a function of the size of *t*.

As we can see, the final value does not appreciably alter based on the size of *t*. The reason for this stability is that the behavior of the learner is dependent on the probability distribution of the data. In case *t* is small, each data point has a larger impact. In case *t* is large, each data point has a smaller impact. But, because the probability distribution is always the same, the learner always ends up with the same value so long as *t* is equal to the number of data points in the learning period. Moreover, if the learner encounters data after having seen *t* amount of data, this data cannot be used to update the probabilities.

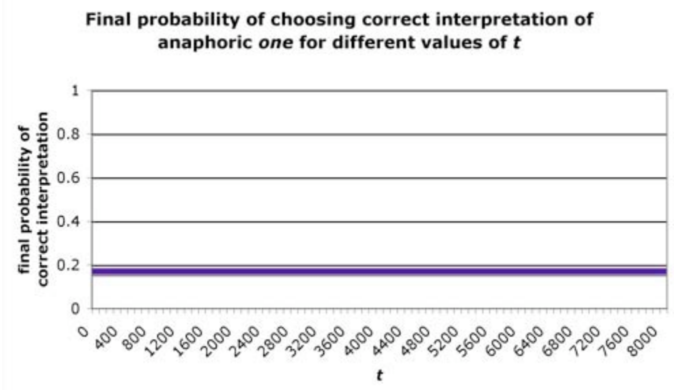


Figure 25. Final probability of the adult interpretation, given different values of *t*. All values are approximately 0.171.

However, suppose the learner encounters *fewer* data points than *t*. For instance, if the *t* for 18-month olds was actually larger than 4017, then the final probability would vary with respect to *t*. Below, I show the final probability for *t* greater than 4017 data points.

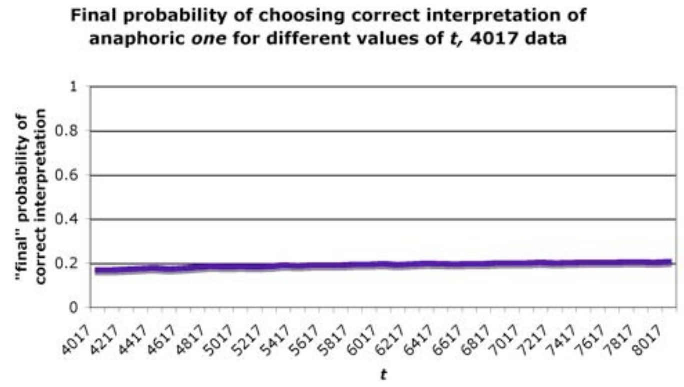


Figure 26. “Final” probability of the adult interpretation, given different values of *t* and a learning period of 4017 data points. Values approach the initial probability of 0.250, reaching 0.206 if *t* is 8017 data points (roughly twice the *t* assumed in the EO Bayesian learner).

Here, we see that the larger *t* is, the less the final probability deviates from the 0.250 initial probability. This is because each data point shifts the probability less, as *t* is larger. What we effectively see is the result of the 4017 data point cut-off (assumed for 18-month olds) not being at the end of the learning period. Thus, the learner (or learner's brain) expects to encounter more data points before settling onto the final probability; the "final" probability at 4017 data points is higher than the ultimate final probability at the end of the learning period. If the learner encounters *t* data points, the final probability will be 0.171, as we saw in figure 25 above.

3.5.7 The Outcome of an EO Bayesian Learner

To summarize, even with conservative estimates of various parameters, the EO Bayesian learner is heavily biased against the preferred adult interpretation of anaphoric *one* in an utterance with two potential antecedents. In fact, the probability of converging on the preferred adult interpretation of anaphoric *one* is quite small (0.171). In short, there is less than a one in five chance of an EO Bayesian learner converging on the correct interpretation for anaphoric *one*.

This result is strikingly different from that reported in R&G, who found overwhelming success for a Bayesian learner. What is the source of this difference? Recall that R&G's model made use of only a subset of the available data and gave priority to semantic data over syntactic data. However, if a Bayesian learner is unconstrained in its data intake, then we would expect that it does not favor one type of data over any other - favoring one type of data over another represents a domain-specific filter.

This EO Bayesian model, in contrast, lacks any domain-specific filter on data intake. It uses all the available data (unambiguous, type I ambiguous, and type II ambiguous) and treats syntactic and semantic data as equally relevant to the learner. As we can see, such an unconstrained domain-general learning procedure on its own fails to converge on the correct interpretation of anaphoric *one* with high probability.

This failure is especially striking because of how generous I was regarding the data available to the EO Bayesian learner and how the learner interpreted that data. Below, I highlight where I was generous and see that revoking that generosity only pushes the final probability of choosing the preferred adult interpretation closer to zero. So, I will conclude that unconstrained (and specifically, unfiltered) Bayesian learning by itself is not sufficient to model human learning or behavior in this domain.

As noted above, there were two places in the construction of the model where I biased the learner towards the correct interpretation of anaphoric *one*. First, I gave a generous interpretation of the available data by providing a liberal estimate of the amount of informative data in the environment. Second, I made conservative assumptions about the learner's understanding of the environment. Even in the face of this generosity, the EO Bayesian learner failed.

In the first case, I was unable to determine a fair estimate of the amount of informative data in the environment – for example, the confidence a learner had in the type I ambiguous data (section 3.5.3.2), the quantity of type I ambiguous data that were informative (section 3.5.3.3), and the quantity of data indicating the nonpreferred adult interpretation (section 3.5.3.3). Consequently, I maximized the size of the informative data set in order to get an upper bound on the probability of converging on the correct interpretation. In what follows, I leave these assumptions as is.

In the second case, however, I show one way in which we can relax the conservative assumptions about the learner's understanding of the environment to make these assumptions more realistic. As we will see, the results reported above represent an upper bound on the probability of converging on the correct interpretation of anaphoric *one* when there are two potential antecedents. Changing the relevant assumptions only decreases this probability further.

The conservative assumption I will examine concerns the value of $p_{\text{N from N'}}$, which is the probability of observing a Noun-only string, given the set of all the N' strings. I previously described the elements of the N' string set as category strings, such as Noun and Adjective Noun. However, if I describe the elements of the N' string set as strings consisting of vocabulary items, such as "bottle" and "red bottle", the probability of observing a Noun-only string is much smaller: it is the number of Noun-only strings divided by the total number of N' strings in the learner's language. The MacArthur CDI (Dale & Fenson, 1996) suggests that 14-16 month olds know about 247 nouns and 49 adjectives. Therefore, the total number of N' strings for an 18-month old learner consists of at least all the nouns and adjective+noun combinations, which is 247+49*247=12350.[18] Using these (still somewhat conservative) estimates, $p_{\text{N from N'}}$ is 0.0201. This is considerably smaller than the previous value of 0.25. Recall that the smaller the value of $p_{\text{N from N'}}$, the more the N' hypothesis is penalized whenever a type II ambiguous data point is encountered.

Using this less generous value of $p_{\text{N from N'}}$ (0.0201, instead of 0.25), the probability of converging on the adult interpretation is the product of the probability of the correct syntactic hypothesis (0.235, 1000 simulations with sd = 0.00316) and the probability of the correct semantic interpretation hypothesis (0.554, 1000 simulations with sd = 0.00358), which is 0.130. On the current, more realistic estimate of the model's parameter, the learner now has less than a one in six chance of converging on the preferred adult interpretation of anaphoric *one* in a situation where there are two potential antecedents for *one.*

3.6 On the Necessity of Domain-Specific Filters on Data Intake

We began our discussion with the observation that a learning theory can be divided into three components: the representational format, the filters on data intake, and the learning procedure. The EO Bayesian learner attempted to solve the problem of anaphoric *one* using a prespecified representational format[19], but no domain-specific filters or learning procedures. In contrast, the model presented by R&G,

[18] Again, this is a conservative estimate since there are still more N' strings from combinations of prepositional phrases as well as adjectives with prepositional phrases, for instance – e.g. "bottle in the corner", "big striped ball behind his back", etc. The effects of recursive modification only exacerbate the problem.

[19] Although our model requires antecedent knowledge of X-bar theoretic structures, it is an independent question whether these are innate or derived from experience.

which also used a prespecified representational format and a domain-general learning procedure, used two domain-specific filters on data intake. This model succeeded. We can now examine (a) whether both of these filters are necessary to converge on the preferred interpretation of anaphoric *one*, and (b) whether we can derive the necessary filters in a principled fashion.

The first filter R&G's learner considers is to use only semantic data. That is, alternative syntactic hypotheses were evaluated only with respect to the predictions they made about the referents of phrases containing anaphoric *one*. These are the semantic consequences of the syntactic hypotheses. However, these hypotheses were not evaluated with respect to the predictions they made about the set of possible strings that would be available as antecedents for anaphoric *one*. So, the syntactic implications of the syntactic hypotheses were not considered. The second filter R&G's learner used was to systematically exclude type II ambiguous data. These are examples in which the antecedent for anaphoric *one* is an NP containing no modifiers (e.g. …*ball*…*one*…).

We can now ask what happens to the EO Bayesian learner if we use these filters, separately and together. First, consider a variant of the EO Bayesian learner that learns only from the semantic consequences of its syntactic hypotheses. In the semantic interpretation domain, that learner maintained two hypotheses: the N'-property hypothesis and the any-property hypothesis. The probabilities of these two hypotheses are updated on the basis of semantic data. Moreover, these hypotheses are linked to the syntactic hypotheses. The N'-property hypothesis is linked to the N' hypothesis (specifically, the exclusive superset of the N'-hypothesis); and, the any-property hypothesis is linked to the N^0-hypothesis. Consequently, by updating the probabilities of the semantic hypotheses, we also update the probabilities of the syntactic hypotheses. If we ignore the syntactic consequences of the hypotheses, then the only way to update the syntactic hypotheses is via the link to the semantic hypothesis space.

If I simulate an EO Bayesian learner that only learns via the semantic analysis of the data, the final probability for $p_{N'}$ and $p_{N'\text{-prop}}$ is 0.810. There is no deviation, since the data points consist of the 10 unambiguous data points, which are maximally informative for the N' and N'-property hypotheses, and the 183 type I ambiguous data points, which I generously assumed were maximally informative for the N' and N'-property hypotheses. Moreover, there are no countervailing data points for the alternative hypotheses (N^0 in the syntax and any-property in the semantics). Thus, the probability for the correct hypotheses is continually increased. Because only data with semantic consequences is considered, the type II ambiguous data is ignored and so its effect on the final probability is nullified. The final probability of converging on the correct interpretation is the product of the two probabilities, which is 0.656. This is a marked improvement over the unfiltered Bayesian learner; the semantics-only filtered Bayesian learner is nearly four times as likely to converge on the preferred adult interpretation of anaphoric *one*. However, this probability is still significantly below the ideal probability of 1.0, which would indicate absolute certainty of choosing the preferred adult interpretation. Analyzing the data only in terms of its semantic interpretation can generate significant improvement, but seems to still fall short of leading the learner to the correct interpretation with high probability.

The second filter that R&G's model used was the exclusion of type II ambiguous data. We can now ask what happens if I follow R&G in excluding this data. This variant of the model will, like the original EO Bayesian learner, take into account both the semantic and syntactic consequences of its hypotheses, but ignore the type II ambiguous data. Note that ignoring the type II ambiguous data is an explicit filter that specifies the exclusion of this type of data, rather than having the exclusion result from a restriction on the semantic interpretation (as in the semantics-only filter we just examined).

To simulate this no-type-II-data filter, I considered only the unambiguous and type I ambiguous data points (193, by my estimate), as in the previous filter. However, both the syntactic data and semantic data was used for updating, thus making use of the link across the two domains and the fact that there are multiple sources of information. When I run the model on this data set, the final probability for the N' hypothesis in the syntax and the N'-property hypothesis in the semantics is 0.930. The product of these two, which represents the probability of converging on the correct interpretation for anaphoric *one* is 0.865. This is again a sharp improvement over the filter-free variant of the model (over 5 times more likely to converge on the correct interpretation). Additionally, the no-type-II-data filter outstrips the semantics-only filter in performance (0.865 probability against 0.656 probability), and is far closer to the ideal probability of 1.0 that indicates certainty for choosing the preferred adult interpretation of anaphoric *one*.

I now consider the consequences of using both of these filters simultaneously. Recall that the effect of the semantics-only filter, which restricted the learner to using only the semantic analysis, was that only semantic data could impact the hypotheses. This results in the type II ambiguous data being excluded from consideration, as it is uninformative with respect to the alternate semantic interpretations since it has only one potential antecedent. The no-type-II-data filter explicitly excludes type II data. So, if the model use these two filters in concert, the result is *the same* as when it used the semantics-only filter alone; the type II ambiguous data is excluded (by the semantics-only filter, due to its lack of semantic consequences, and by the no-type-II-data filter explicitly) and only semantic data can impact the probabilities associated with the hypotheses (due to the semantics-only filter). Thus, the resulting probabilities for the N' hypothesis and N'-property hypothesis are 0.810 and the probability of the preferred adult interpretation of anaphoric *one* is 0.656. Since using both filters yields an identical result to using the semantics-only filter alone, the benefit gained from using the no-type-II-filter is lost. It is therefore in the interest of the learner to apply only the no-type-II-filter. That is, the learner should ignore type II ambiguous data, but still use both syntactic and semantics data equally to update the hypothesis spaces.

To summarize, the EO Bayesian learner shows us that a learner not equipped with domain-specific filters on data intake cannot converge on the correct interpretation for anaphoric *one*. Figure 27 displays the learning trajectories and outcomes for the full set of simulations: no filter, semantics-only filter, no-type-II-data filter, both filters. As we can see, using the no-type-II-data filter by itself yields the highest probability for the correct interpretation. Moreover, the efficacy of this filter is negated when used with the semantics-only filter. In other words, the ideal learner must use both syntactic and semantic evidence, but be restricted in which sentences it takes as opportunities to learn from.

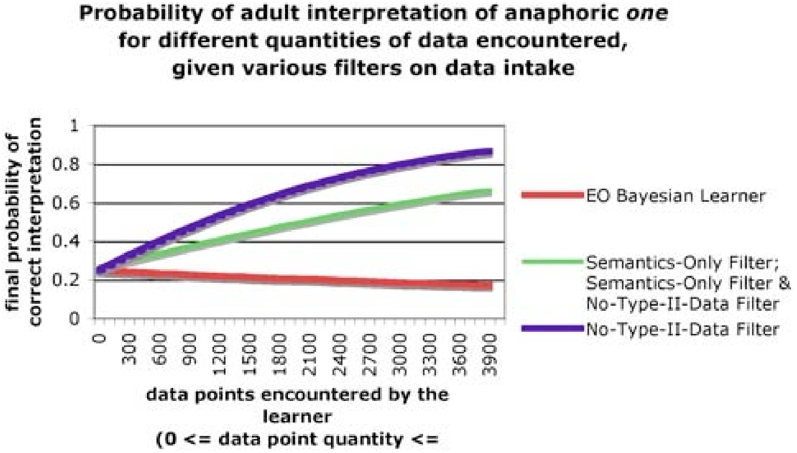


Figure 27. The Bayesian Learner's trajectory as a function of the amount of data encountered: no filters, semantics-only filter, no-type-II-data filter, and both semantics-only filter and no-type-II-data filter.

3.7 Deriving the Necessary Domain-Specific Filter

The necessity of a filter on data intake now raises an important question. Where does this filter come from? It seems fairly obvious that the learner cannot come equipped with a filter that says "ignore type II ambiguous data" without some procedure for identifying this data. What we really want to know is whether there is a principled way to derive the existence of this filter. Specifically, we want the filter to ignore type II ambiguous data to be a consequence of some other principled learning strategy.

Suppose there is a general principle that learning occurs only in cases of uncertainty, because it is only in cases of uncertainty that information is conveyed (Shannon 1948; cf. Gallistel 2001). The learning algorithm therefore engages only when there is uncertainty about the identity of the antecedent.

One suggestion would be to call on the semantics-only filter, arguing that interpreting anaphoric *one* is simply a semantic problem. This could be termed a semantocentric approach to learning, and so the syntactic implications are irrelevant for learning. The result of this strategy would be that the learner only uses the semantic consequences of the data to update the hypotheses. As we saw in the previous section, this would rule out type II ambiguous data (with a single string as potential antecedent, such as *ball*), because such data has only one semantic interpretation available (any-property)– thus, there is no uncertainty. However, as we also saw in the previous section, this causes the learner to lose the useful effect that the *syntactic* data can have. Specifically, if only semantic data are used, the benefit gained from having linked domains is lost. The learner uses only semantic data to update the both hypothesis spaces; the learner does not also use the syntactic aspect of the data to update both hypothesis spaces. This leads to a lower probability of converging on the adult interpretation of anaphoric *one*.

Another suggestion is that the learner takes a syntactocentric approach, and the problem the learner faces is solely to do with the string that is the antecedent of anaphoric *one*. The only influence semantic interpretation data has is as a reflection of various syntactic hypotheses that are entertained. Suppose that the learner comes equipped with a constraint against anaphora to X^0 categories (Baker, 1979; Hornstein & Lightfoot, 1981) or is able to have derived it previously using a syntactocentric filter on the available data (Foraker et al, in press). The syntactic hypothesis space is reduced to a single hypothesis: *one* = N'. In this situation, the learner needs only to solve a different problem in the syntax domain: namely, which N' is the appropriate antecedent in cases in which there are multiple N's available.

For example, if the learner hears "Here's a red ball. Give me another one, please," there are two N's available, *red ball* and *ball*. These two different antecedents have different semantic interpretations: *red ball* is restricted to red balls whereas *ball* is not. In other words, the N'-property hypothesis is linked to the larger N' *red ball*, whereas the any-property hypothesis is linked to the smaller N' *ball*. Choosing the appropriate antecedent can be achieved using the update functions described for the EO Bayesian learner.

Now, in cases in which there is only one N' available (as in type II ambiguous data), there are no choices to be made in finding an antecedent. That is, if the learner hears, "Here's a ball. Give me another one, please," the only possible antecedent is the N' *ball*. Consequently, the learner has no uncertainty about the meaning of the expression and so does not invoke the learning algorithm.

This last point is critical for motivating the learner's choice to ignore type II ambiguous data. As noted above, having a range of available antecedents causes uncertainty about the antecedent. It is this uncertainty that triggers the learning algorithm. It is important to see at this point that this syntactocentric approach requires the learner to be concerned not with the category of the antecedent (N' vs. N^0), but rather the identity of the antecedent when there are two or more N's to choose from. However, allowing the learner to view this as a problem of which syntactic antecedent to choose rather then merely as a problem of interpretation causes the learner to use the syntactic aspect of the data as well, which we found was crucial for a more successful learner.

3.8 Future Directions

Learning anaphoric *one* is a case study that can be mined further still. For example, we can consider if learning success is possible in a hypothesis space that contains more than two hypotheses in a subset-superset relationship. Does the learner only consider two overlapping hypotheses at a time (small N' *ball* vs. larger N' *red ball*), or can the learner achieve success when, say, three hypothesis are considered concurrently (small N' *ball*, larger N' *red ball*, even larger N' *big red ball*)?

Moreover, we can open up the current hypothesis space containing only two possible N's even more if we allow the learner to entertain syntactic hypotheses involving antecedents containing covert modifiers. Suppose, for example, that the learner hears, "Look, a bottle! Oh, and it's red! Jack doesn't have one like that." Suppose also that Jack has a non-red bottle, so it is clear that *one* refers to a red bottle in the world. The difficulty for the learner is that the antecedent of *one* in the available utterances is overtly *bottle*, but it is implicitly *red bottle* (as the bottle Jack doesn't have is a red bottle). Yet, *red bottle* does not appear overtly in the data. The learner might then need to entertain a hypothesis where the antecedent contains a covert modifier that corresponds to the property the referent in the world has, e.g. (*red*) *bottle* when the referent in the world is a red bottle. This would alter how the learner updates the probabilities associated with each hypothesis when considering information from both the potential syntactic antecedents and semantic referents in the world for anaphoric *one* data points.

I do note that before pursuing this it is worthwhile to determine via standard experimental techniques, such as those used by LWF (2003), how real learners interpret a data point of this kind. If they do interpret *one* as referring to a red bottle in the example above (and so having a linguistic antecedent of *red bottle*, even though it is not explicit in the utterance), then the question of how to expand the learner's syntactic and semantic hypothesis spaces appropriately becomes particularly relevant.

In addition, I have defined the hypothesis spaces by the number of data types that are compatible with each hypothesis (e.g. Noun, Adjective Noun, etc.). But we might also include frequency of data type, especially when considering the relative size of one hypothesis space against another. For instance, suppose the N^0 hypothesis space consists of data types{Noun} and the N' hypothesis space consists of data types {Noun, Adjective Noun}. The N' hypothesis space is twice as big as the N^0 hypothesis, under this definition. But suppose the learner has encountered 9 examples of Nouns and 1 example of Adjective Noun. Then the N' hypothesis space is only 1/10 larger than the N^0 hypothesis space, given the learner's current experience. This then influences the updating that occurs when encountering an ambiguous data point (Noun). The relative size of the hypothesis spaces alters over time, as the learner encounters more examples from the input. So, the impact of ambiguous data likewise alters over time. Under these conditions, is acquisition success possible without filtering the data intake? This is certainly a question worth exploring.

3.9 Conclusion

The case of anaphoric *one* demonstrates the interplay between domain-specificity and domain-generality in learning. What we have seen here is that a domain-general procedure can be successful, but crucially only when paired with domain-specific filters on data intake. Moreover, I have suggested that the particular domain-specific filter that yields the best result can plausibly be derived from a domain-specific constraint on representation (either innately specified or derived via a syntactocentric analysis).

In addition, I have tried to highlight the consequences associated with the existence of multiple, connected levels of representation in language. Because the levels of representation are linked to each other, conclusions drawn by the learner in one domain also ramify in other domains. When the learner used both syntactic and semantic information with no filters, the result was very poor learning. When the learner used both syntactic and semantic information , in concert with the no-type-II-data filter, the result was very good learning. However, when I disconnected the two domains, as when the learner learned only from semantic data, the result was learning that was not as good (though still much better than no filtering of the data at all). This was due to some of the available information – the syntactic implications of the syntactic hypotheses – being ignored. Thus, the connection between domains allows multiple analyses across domains of a single data point to each have an effect. This, in turn, will magnify the effect of a given data point, thus increasing the amount of information that can be salvaged by the learner. This lesson should be generalized to learning in any situation involving multiple linked levels of representation.

Finally, it is important to recognize that I have simulated learning only for one very specific case of grammar acquisition. However, the inherent semantic compositionality of syntactic representations provides a severe hurdle for Bayesian learning techniques that are biased towards the most restrictive hypothesis. As I have noted, as the syntactic structure grows, the set of referents in the semantics shrinks. Consequently, the most restrictive hypothesis in the syntax corresponds to the least restrictive hypothesis in the semantic interpretation, and vice versa. This makes it impossible to define a "most restrictive hypothesis" across both domains.

The existence of multiple, linked levels of representation in language, and presumably elsewhere in cognition, has important consequences for learning. A link between domains can amplify the positive effects that come from using data from multiple sources. Nonetheless, this link can structure the data in such a way as to nullify the essential advantage of unconstrained Bayesian learning techniques.

Chapter 4: The Case of Old English Word Order

4.1 Filters on Data Intake for Syntactic Learning

The phenomenon I examine in this chapter is an instance of syntactic learning, specifically the alternation between Object-Verb (OV) and Verb-Object (VO) order in Old English. This case is another example where the learner has two hypotheses under consideration. However, unlike the case of anaphoric *one*, the final state for *adults* in Old English is argued to be probabilistically distributed between the two hypotheses (Pintzuk, 2002; Kroch & Taylor, 1997; Bock & Kroch, 1989). Evidence for this mixed adult state comes from texts in which both alternates are exhibited by a single author. This is in contrast to final state where only one hypothesis is accessed (i.e. only one structural rule used) by adults.

The hypothesis space for Old English OV/VO order consists of two hypotheses that overlap, but do not have a subset-superset relation. Both the OV and VO hypotheses have data that will be unambiguous. In addition, there is a quantity of data that is ambiguous between the two hypotheses since it can be analyzed successfully given either hypothesis. The updating procedure is based off the one described in the mathematical framework in chapter 2. I then use this definition of the hypothesis space and the updating procedure to investigate two filters on data intake proposed for syntactic learning.

The two filters in question bias learners away from potentially misleading ambiguous data in the input, both stemming from a presumed preference for "simple" data (Dresher, 1999; Lightfoot, 1999, 1991; Fodor, 1998a). These filters use a structurally-based notion of simplicity. The first claims that children learn only from unambiguous data (Dresher, 1999; Lightfoot, 1999; Fodor, 1998a), and consequently do not activate the update algorithm whenever data is 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, the update algorithm again is not activated and these data are effectively ignored by the learner.

These filters are motivated by the perceived informativity and ease of comprehensibility of the relevant data. As we saw in the previous chapter, an unambiguous data point allows the learner to be maximally confident in whichever hypothesis the data point signals. So, the most probability is shifted when the learner encounters an unambiguous data point. We can view this as unambiguous data points being the most informative data points available to the learner. For simple clauses, it has been claimed that children might restrict their attention to simple, subparts of utterances (Morgan, 1986), perhaps because of general cognitive restrictions on the complexity of data that they can handle. So, matrix clauses, being "simpler", are arguably easier for learners to extract information from.

Nonetheless, filtering the data is not without its drawbacks. The filters proposed above will radically truncate the data intake set. It is well known that sparse data can inhibit a probabilistic model's ability to converge on a solution. Thus, we must determine if the subset of data circumscribed by these two filters can still allow learning to succeed, even if the subset is significantly smaller than the input data set.

In Old English, as we have already noted, the adult state is a probabilistic distribution between the two hypotheses, OV and VO word order. 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?

At this point, we can 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. (YCOE Corpus, Taylor et al., 2003; PPCME2 Corpus, Kroch & Taylor, 2000). It has been proposed that certain types of change (such as the shift in Old English) result from a misalignment of the child's hypothesis and the adult's analysis of the same data (Lightfoot, 1999; 1991). In other words, language change in this case results from *imperfect* learning of a very particular kind.

Specifically, 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 must hypothesize exactly the right amount of difference between the learner's and adult's probability distributions.

Therefore, "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. We will find that the amount of misconvergence depends greatly on how the input is filtered during learning. Thus, we can test proposals about data filtering by using models of language change.

It is important to note the correlation between successful learning and imperfect learning for certain cases of language change. Often, language learning research in synchronic cases may focus so much on the learner's ability to reach the target adult state that we may overlook the fact that perfect learning will not necessarily lead to success in diachronic cases. This is because perfect learning would entail no change over time. This then creates a certain tension on the demands of a successful learning model – it must be good enough that learners can communicate effectively with the remainder of the population, but not so good that language change is impossible. So, using successful language change as a metric for successful language learning attempts to keep this second point in mind.

We will find, perhaps surprisingly, that the two proposed filters on data intake are crucial for a successful model of Old English language change that describes a population which begins strongly OV-biased at 1000 A.D. and ends strongly VO-biased at 1200 A.D. Without these filters, 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. This supports the existence of these two filters on data intake during the normal course of syntactic learning.

The chapter proceeds as follows. First, I will discuss the two filtering proposals in detail. Then, I will examine the available information on the language change in Old English. After that, I will discuss the model of language learning and language change that I will use. Finally, I will present the modeling results and discuss their implications for language learning.

4.2 Restricting the Data Intake

4.2.1 Unambiguous Data

4.2.1.1 Unambiguous Data for OV and VO Word Order

Unambiguous data is defined within a hypothesis space of opposing analyses for a certain piece of linguistic structure, such as OV or VO word order. Ambiguity is often faced by a child choosing the correct grammar for his or her language. Let's consider a simple example. The child has to decide whether the stream of encountered speech belongs to a VO (Verb before Objects) language requiring rules like (1) or to an OV (Objects before Verbs) language requiring rules like (2).

(1) VO rule set examples
(a) VP → V NP PP (b) VP → V NP

(2) OV rule set examples
(a) VP → NP PP V (b) VP → NP V

Modern English chooses the VO rule set (1). Modern Dutch and German choose the OV rule set, which includes those in (2). However, modern Dutch and German also generate strings that are compatible with some of the rules in set (1), such as in (3) below:

(3) Ich$_{\text{Subj}}$ sehe$_{\text{TensedVerb}}$ [den Fuchs]$_{\text{Obj}}$
I *see* *the fox*
'I see the fox.'

This example demonstrates an option available in modern Dutch and German which moves the tensed Verb of the matrix clause to the "second" phrasal position in the matrix clause, known as V2 movement (Lightfoot, 1999; Kroch & Taylor, 1997; among many others). The tensed Verb *sehe* moves from its original position (after *den Fuchs*) to the second phrasal position in the sentence, and some other phrase (*Ich*) moves to the first phrasal position, as in (4).

(4) Ich$_{\text{Subj}}$ sehe$_{\text{TensedVerb}}$ t_{Subj} [den Fuchs]$_{\text{Obj}}$ $t_{TensedVerb}$.
I *see* *the fox*
'I see the fox.'

Given the example in (3), one might reasonably wonder why we posit the analysis in (4) instead of simply assuming that modern German (and Dutch) word order is VO. The reason is that VO order does not appear in matrix clauses across the board. Languages like modern Dutch and German use VO order only for tensed Verbs in matrix clauses. Non-tensed Verbs in matrix clauses and all Verbs in embedded clauses obey OV order and appear after the Object. This forces us to assume a basic OV word order with an additional operation that moves the tensed Verb in matrix clauses.

In the bold part of (5a), we see the basic OV order appearing in the embedded clause as *den Fuchs sehen kann* (Object Non-TensedVerb TensedVerb). In (5b), the non-tensed Verb *sehen* appears in the matrix clause after the Object *den Fuchs*, again displaying the OV order. The V2 rule moves the tensed modal *kann* to the second phrasal position, and the Subject *Ich* moves to the first phrasal position.

(5a) Ich$_{\text{Subj}}$ denke$_{\text{TensedVerb}}$, das ich **[den Fuchs]$_{\text{Obj}}$**
I think that I the fox

sehen$_{\text{Non-TensedVerb}}$ kann$_{\text{TensedVerb}}$
see can

‘I think that I can see the fox.’

(5b) Ich$_{\text{Subj}}$ kann$_{\text{TensedVerb}}$ t_{Subj} [den Fuchs]$_{\text{Obj}}$ sehen$_{\text{Non-TensedVerb}}$ $t_{TensedVerb}$
I can the fox see
‘I can see the fox.’

At the beginning of language learning however, the child has not set the word order parameter for the language. Therefore, both the OV and VO hypotheses are available with some probability. The matrix clause *Ich sehe den Fuchs* can be covered by both hypotheses. The OV hypothesis can use the analysis described in (4), matrix OV order with the V2 movement rule; the VO hypothesis can use the analysis in (6), matrix clause VO order without the V2 movement rule (which is the analysis used for modern English).

(6) Ich$_{\text{Subj}}$ sehe$_{\text{TensedVerb}}$ [den Fuchs]$_{\text{Obj}}$.
I see the fox
‘I see the fox.’

Data points like (3) are therefore ambiguous between the two hypotheses under consideration. A proposal to filter data intake down to the unambiguous data points would cause the learner not to activate the update procedure when encountering ambiguous data points.[20] Instead, the learner uses only data points perceived as unambiguous. Examples of perceived unambiguous data are in (5) above. In (5a), if the child uses embedded clause data as intake, then the presence of

[20] Otherwise, the learner would require some strategy for how to update the probabilities when encountering ambiguous data, as we saw in the previous chapter.

the Verbs (both tensed *kann* and non-tensed *sehen*) after the Object would signal that the VO hypothesis is correct. In (5b), the presence of the non-tensed Verb *sehen* after the Object again implicates the OV hypothesis since that order would not be generated by a VO system.

4.2.2.2 Identifying Unambiguous Data

If we believe that children filter their intake for syntactic learning down to unambiguous data, it is important to provide a plausible method for identifying unambiguous data. Two methods have been proposed to identify unambiguous data: the domain-specific knowledge of cues (Dresher, 1999; Lightfoot, 1999) and the domain-specific procedure of parsing (Fodor, 1998; Sakas & Fodor, 2001).

A cue for identifying unambiguous data is defined as a specific configuration in the surface structure of the data point that signals one parameter value (hypothesis) is correct. The knowledge of what a cue for a given parameter value looks like is often presumed to already be available to the learner (Dresher, 1999; Lightfoot, 1999), whether innately specified or derived through some other knowledge. A cue for OV/VO word order proposed by Lightfoot (1999) is described in (7).

(7) The Object is adjacent to the Verb (on the appropriate side) and the Verb is *not* in the second phrasal position.

This is considered a cue because the V2 movement rule deriving a VO order from an underlying OV order only allows a single phrasal constituent to come before the Verb. If the Verb is preceded by more than one phrasal constituent, then its position is not the result of V2 movement.[21] The form of this cue could be an underspecified piece of sentence structure (figure 28 below) or simply a linear pattern retrievable from the observable data (8). Both are representations of the domain-specific knowledge that a cue describes.

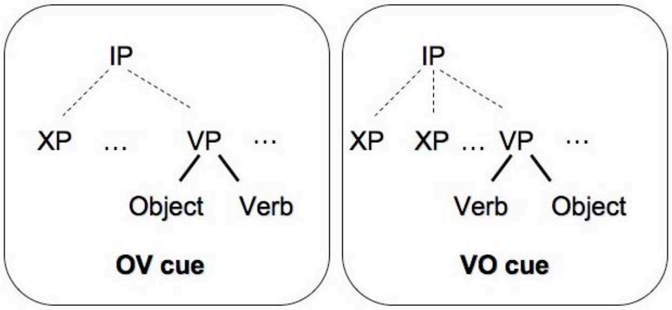


Figure 28. Underspecified pieces of sentence structure that could be the learner's representation of a cue for OV vs. VO word order, as described by Lightfoot (1999).

[21] Note that this is the learner's *perception* of the data, given a restricted knowledge base. The adult grammar, in actuality, may contain other grammatical rules that allow V2 movement to create a clause with the Verb in the third position. Thus, the learner may perceive data as "unambiguous" that is ambiguous when a fuller range of grammatical rules is considered.

(8) Linear patterns that could be the learner's representation of a cue for OV vs. VO word order.

(a) OV cue: $[\ ]_{XP}$ … Object Verb …
(b) VO cue: $[\ ]_{XP1}$ $[\ ]_{XP2}$ … Verb Object …

To identify unambiguous data, the learner matches the data point (or relevant piece of the data point) to the cue. Example sentences that would match these cues are in (9).

(9a) Matching OV cue: Subject Object Verb.
Ich denke, das *ich den Fuchs sehe.*
(XP = Subject, … = *null*)
(9b) Matching VO cue: Adverb Subject Verb Object.
Yesterday, I saw a dragon.
(XP1 = Adverb, XP2 = Subject, … = *null*)

The cues method gives sentences like these privileged status, and such sentences are viewed as unambiguous evidence for the associated parameter value, OV or VO.

An alternative approach is to use the learner's natural language comprehension processes to discover if a data point should be considered unambiguous for OV/VO order (Fodor, 1998b; Sakas & Fodor, 2001). The learner assigns possible structures to (or *parses*) the datum with all values of the *relevant* parameter set (in this example, the relevant parameter set *PS* = {OV/VO, +V2/-V2}).[22] If only one value of a parameter (e.g. OV) will allow a successful parse of the entire data point, then that data point is classified as unambiguous for that value of that parameter. This procedure is shown in (10).

(10) Parsing to identify unambiguous data for basic word order using the set of parameter values PS = {OV/VO, +V2/-V2}

(a) Data point: *Subject Object Verb.*
Sets of values from PS that will lead to a successful parse of the data =
{OV, -V2}
In this case, the only combination of values that will allow a successful parse is OV and –V2. Therefore, given this set of relevant parameter values, this data point is unambiguous for both OV and –V2.

(b) Data point: *Subject TensedVerb NonTensedVerb Object.*
Sets of values from PS that will lead to a successful parse of the data =
{VO, +V2}, {VO, -V2}
In this case, two combinations of values will allow a successful parse of the data point, and both use the VO value (and neither use the OV value). Either value of the V2 parameter can be used in combination with the VO value, however. Therefore, given this set of relevant parameter values, this data point

[22] Note that the relevant parameter set for the learner may be (and likely is) a subset of the entire adult parameter set.

is unambiguous for VO only.

(c) Data point: *Subject Verb Object.*
Sets of values from PS that will lead to a successful parse of the data =
{OV, +V2}, {VO, -V2}, {VO, +V2}
In this case, three combinations of values will allow a successful parse of the data point. Importantly, neither parameter value for either parameter is crucial for parsing success. There is at least one combination that uses the OV value, at least one that uses the VO value, at least one that uses the +V2 value, and at least one that uses the –V2 value. Therefore, given this set of relevant parameter value, this data point is *not* unambiguous for any values of any parameters.

We will return in the next chapter to the discussion of the benefits and drawbacks of each method that the learner could use in identifying unambiguous data. For the case of Old English OV/VO order discussed in this chapter, both methods will identify the same set of utterances as unambiguous data, provided the relevant parameter set for parsing is restricted as described above.[23]

4.2.2.3 Unambiguous Data Summary

The unambiguous data filter reflects a very simple idea: the child learns only from the data perceived as "clean", instead of guessing about data perceived as "unreliable". If the child is using cues, clean data are identified by the specific rubric of the cue. If the child is using parsing, clean data are identified by having only one parameter value that yields successful parsing. For both methods, it is important to note that a data point is unambiguous relative to a given parameter. A data point unambiguous for parameter P1 may not be unambiguous for another parameter P2. For instance, as we saw in (10b), a data point can be unambiguous for VO order while being ambiguous for the V2 movement operation.

In addition, an unambiguous filter reduces the set of data a child can learn from (since some data in the input are classified as unambiguous). It is therefore quite important that there be enough data left in the child's intake to learn from. If the data perceived as unambiguous appear in sufficient quantity in the input, the learner will converge on the "correct" probability distribution for that parameter. Otherwise, the individual learner within the population will not be able to converge on the correct probability distribution, and will instead remain near the initial probability distribution. Once individuals are unable to converge on the correct probability distribution, language change in the population as a whole will grind to a halt. Thus, it is critical for the feasibility of an unambiguous data filter that the unambiguous data not be too sparse in the input.

[23] Specifically, the relevant parameter set for parsing should not include operations that can influence the position of the Object with respect to the Verb, such as Heavy Noun Phrase shift which will move the Object to a position following the Verb if the Object is phonologically "heavy enough". If the parameter set *did* include operations like this, many more data points would be considered ambiguous and therefore unusable for a learner employing an unambiguous data filter.

4.2.3 Simple Clauses

The potential problem of data sparseness becomes worse when we add a proposal to learn from data in simple clauses only: the “degree-0” learning filter of Lightfoot (1991). Degree refers to the level of embeddedness. I adopt Lightfoot’s terminology “degree-0” to refer to matrix clauses and “degree-1” to refer to embedded clauses.[24] This filter is motivated by a claim that it lessens the cognitive load of the learner; children use only structural information that spans a single matrix clause and at most a complementizer in the embedded clause.[25] A learner using this filter would not use data such as (5a) as evidence for the OV order of German, since the useful structural information signaling OV order is in the embedded clause. Nonetheless, examples such as (5b) that contain non-tensed Verbs adjacent to the Object in the matrix clause are still in the degree-0 learner’s intake.

4.2.4 The Influence of Input Filtering on Old English Language Change

Potential data sparseness aside, filtering of the input can go a long way toward explaining how changes to a language’s structure can spread fairly rapidly through a population. Filtering requires learners to learn only from a specific subpart of the observable data. If that subpart changes (perhaps due to external factors) so that it does not accurately reflect the adult probability distribution for the language as a whole, then children will “mislearn” the adult probability distribution. These children subsequently contribute observable data to the next generation of children, who will subsequently “mislearn” the previous children’s “mislearned” probability distributions. This continues, spreading through the exponentially growing population[26], until the population as a whole has shifted its probability distribution dramatically.

The loss of a strongly OV distribution in Old English is an especially interesting language change because the degree-0 unambiguous data distribution of the two word orders appears to be significantly different from the average adult’s probability distribution for the language as a whole. The V2 rule’s restriction to matrix clauses means that while the distribution of clauses in the matrix is mixed between VO and OV order, Old English (before the change) is strongly OV in

[24] Lightfoot’s work follows Wexler & Culicover (1980) and Morgan (1986), who argue for less restrictive constraints on the learning domain.

[25] Note that this motivation wouldn’t necessarily hold for head-final languages like Japanese where the matrix clause can be split into two parts by an embedded clause: $\text{Subject}_{\text{Main}}$…$\text{Subject}_{\text{Embedded}}$ …$\text{Object}_{\text{Embedded}}$ $\text{Verb}_{\text{Embedded}}$…$\text{Object}_{\text{Main}}$ $\text{Verb}_{\text{Main}}$. A degree-0 learner would need to track information spanning the embedded clause. A learner with the cognitive resources to do that would most likely also have the cognitive resources to track the information in the embedded clause. So, a degree-0 learner that is motivated by a limit on cognitive resources and who must learn a head-final language might be redefined as one using the information in the portion of the degree-0 clause that is adjacent, i.e. not split by any embedded clause material.

[26] Populations canonically grow at an exponential rate, with the current set of new population members typically outnumbering the previous set of new population members. The exact amount that the current set of new members outnumbers the previous set of new members is described by the population growth coefficient, a constant value specific to a given population.

embedded clauses (see table 4.1 in section 4.4.1.2). This is a case where unambiguous data and degree-0 data filters on data intake should create a mismatch between the adult's underlying probability distribution and the probability distribution the child converges on.

Since we have historical records allowing us to calculate the rate of change from OV to VO, I model the effect of filtering by restricting my model to learn from simple unambiguous structures in the quantities found in the historical record at the beginning of the transformation of Old English from OV to VO. The model will then create a set of successive generations, each diverging from the initial distribution to a designated extent; this is the rate of change. Then, I can calculate the effect of these two filters on the rate of change in the model, and compare it to the actual rate calculated from the distribution of data found at various periods during this transformation in the actual historical record.

I do this in two steps. First, I ask if a population whose learners filter their input down to degree-0 unambiguous data is able to follow the historically attested trajectory. Then I ask whether a model that uses additional data (ambiguous or embedded or both) during learning could also produce the observed historical patterns in the simulated population. This provides us with the evidence we need to determine if children should use these filters during language learning.

4.3 Old English

4.3.1 OV and VO word order in Old English

Between 1000 A.D. and 1150 A.D., the distribution in the Old English population consisted of mostly OV order utterances (11a) while the distribution in the population at 1200 A.D. consisted of mostly VO order utterances (11b) (YCOE Corpus, Taylor et al., 2003; PPCME2 Corpus, Kroch & Taylor, 2000).

(11a) he$_{\text{Subj}}$ Gode$_{\text{Obj}}$ þancode$_{\text{TensedVerb}}$
he God thanked
'He thanked God'
(*Beowulf,* 625, ~1100 A.D.)

(11b) & [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ awecð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
"And with his stem, he awakened the dead to life."
(*James the Greater,* 30.31, ~1150 A.D.)

4.3.2 Unambiguous Data

4.3.2.1 Unambiguous OV

Unambiguous data for OV word order correlate with observable data of the following types in Old English: (12a) the tensed Verb appears at the end of the clause or (12b) the non-tensed Verb remains in the post-Object position, while the tensed auxiliary moves.

(12a) he$_{\text{Subj}}$ hyne$_{\text{Obj}}$ gebidde$_{\text{TensedVerb}}$
He him may-pray
'He may pray (to) him'
(*Ælfric's Letter to Wulfsige,* 87.107, ~1075 A.D.)

(12b) we$_{\text{Subj}}$ sculen$_{\text{TensedVerb}}$ [ure yfele þeawes]$_{\text{Obj}}$ forlæten$_{\text{Non-TensedVerb}}$
we should our evil practices abandon
'We should abandon our evil practices.'
(*Alcuin's De Virtutibus et Vitiis,* 70.52, ~1150 A.D.)

4.3.2.2 Unambiguous VO

A reasonable assumption might be that unambiguous VO data should be the counterpart of unambiguous OV data in form. Specifically, one might assume that since *Subject Object TensedVerb* is unambiguous OV data, *Subject TensedVerb Object* should then be unambiguous VO data. However, recall the V2 movement rule, which moves the tensed Verb to the second phrasal position of the clause. As we will see below, when this movement rule is taken into account, sentences of the form *Subject TensedVerb Object* cannot be perceived as unambiguous VO data.

4.3.2.2.1 V2 Interference

Assuming V2, a simple *Subject TensedVerb Object* utterance could be parsed with either the OV (with V2 movement) or the VO order parameter value (with or without V2 movement). Example (13) shows this: the tensed Verb *clænsað* could begin in sentence final position (OV order) and move to the second position (13a), or it could be generated in this position all along (VO order) (13b).

(13a) heo$_{\text{Subj}}$ clænsað$_{\text{TensedVerb}}$ t_{Subj} [þa sawle þæs rædendan]$_{\text{Obj}}$ $t_{TensedVerb}$
they purified the souls [the advising]-Gen

(13b) heo$_{\text{Subj}}$ clænsað$_{\text{TensedVerb}}$ [þa sawle þæs rædendan]$_{\text{Obj}}$
they purified the souls [the-advising]-Gen
'They purified the souls of the advising ones.'
(*Alcuin's De Virtutibus et Vitiis*, 83.59, ~1150 A.D.)

Because of V2 movement, unambiguous VO data in matrix clauses appears as the examples in (14): there is either (a) more than one phrase to the left of the Verb ([*mid his stefne*]$_{\text{PP}}$ *he*$_{\text{Subj}}$), ruling out a V2 analysis, or (b) some sub-piece of the verbal complex (*up*$_{\text{Verb-Marker}}$) immediately preceding the Object.

(14a) & [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ awecð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
'And with his stem, he awakened the dead to life.'
(*James the Greater,* 30.31, ~1150 A.D.)

(14b) þa$_{\textbf{Adv}}$ ahof$_{\text{TensedVerb}}$ Paulus$_{\text{Subj}}$ up$_{\text{Verb-Marker}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul up his head
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

4.3.2.2.2 Verb-Markers

I will term sub-pieces of the verbal complex "Verb-Markers". A Verb-Marker is a word that is semantically associated with a Verb, such as a particle ('up', 'out'), a non-tensed complement to tensed Verbs, a closed-class adverbial ('never'), or a negative ('not') (Lightfoot, 1991). Under the assumption that the learner believes all Verb-like words should be adjacent to each other (Lightfoot, 1991), a Verb-Marker can be used to determine the original position of the Verb. For (14b), the Verb-Marker *up* indicates the position where the tensed Verb originated before V2 movement; since the Verb-Marker precedes the Object, the original position of the Verb is assumed to be in front of the Object as well. So, this utterance type is perceived as unambiguous data for VO order. Examples of utterances with Verb-Markers are in (15) below (Verb-Markers are in bold): the particle *up* is a Verb-Marker in (15a) and the non-tensed Verb *gewyrecean* is a Verb-Marker in (15b).

(15a) þa$_{\textbf{Adv}}$ ahof$_{\textbf{TensedVerb}}$ Paulus$_{\textbf{Subj}}$ **up**$_{\textbf{Particle}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul ***up*** *his head*
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

(15b) Swa$_{\textbf{Adv}}$ sceal$_{\textbf{TensedVerb}}$ [geong guma]$_{\textbf{Subj}}$ gode$_{\textbf{Obj}}$
Thus shall young men good-things
gewyrecean$_{\textbf{Non-TensedVerb}}$
perform
'Thus shall young men perform good things.'
(*Beowulf,* 20, ~1100 A.D.)

Interestingly, Old English Verb-Markers (unlike their modern Dutch and German counterparts) were *unreliable* as a marker of the Verb's original position. In many cases (such as the negative *ne* in (15c) below), the Verb-Marker would not remain adjacent to the Object. If there were no other Verb-Markers adjacent to the Object, then no indication of the Verb's initial position remained and the utterance could be interpreted as ambiguous between the OV or VO order hypotheses. In (15c), the adverbial *næfre* remains adjacent to the Object, and so this data point would be perceived as unambiguous for VO order.

(15c)	**ne**$_{\textbf{Negative}}$	geseah$_{\text{TensedVerb}}$	ic$_{\text{Subj}}$	**næfre**$_{\textbf{Adverbial}}$	[ða burh]$_{\text{Obj}}$
	NEG	*saw*	*I*	***never***	*the city*

‘Never did I see the city.’
(Ælfric, *Homilies.* I.572.3, between 900 and 1000 A.D.)

4.3.3 Causes of Language Change

4.3.3.1 The Effect of the Unambiguous Data Distribution

As we have just seen, matrix clause cues (such as the location of a Verb-Marker with respect to the Object) can be unreliable. This causes data that potentially could have been perceived as unambiguous to be perceived as ambiguous. Thus, a learner using an unambiguous data filter would potentially encounter a distribution of OV and VO data points that is different from the distribution the adult speakers of the population used to generate the entire data set. In short, the learner’s intake can have a different distribution than that of the available input. This difference in the intake can cause successive generations of Old English children to have different OV/VO probability distributions than their predecessors. The Old English population would then shift to a strongly VO-biased distribution because of what the learners’ intake consists of. I will formally model this intuition by using actual quantitative data from the relevant historical periods coupled with an explicit probabilistic model.

4.3.3.2 A Concern About Other Causes of Language Change

Before we examine the details of the model, I should address a concern about the cause of this particular language change in Old English. I have assumed, based on Lightfoot’s (1991) claim, that language learning (an internal factor) is the instigator of the shift from a strongly OV-biased distribution to a strongly VO-biased distribution. However, one might wonder if external factors could have played a more significant role in this change.

I consider two potential external factors below: Scandinavian influence and Norman influence. We will see that neither factor *by itself* could have caused the change in Old English from a strongly OV-biased distribution to a strongly VO-biased distribution. However, it is still possible that the correct combination and influence of external factors could have produced the recorded historical change, even in the absence of the imperfect learning approach advocated by Lightfoot (1991) and adopted here. The contribution of the present work would then be to demonstrate how it is not *necessary* to have external factors in order to cause abrupt change at the population-level in such a limited timeframe.

4.3.3.2.1 Scandinavian Influence

Scandinavian influence before 1000 A.D. is claimed to have caused Old English Verb-Markers to become unreliable (Kroch & Taylor, 1997). Old Norse, the language spoken by the Scandinavians, used VO order and therefore introduced variability into the OV ordered Old English. Is it possible that continued Scandinavian influence *alone* caused the sharp change in the OV/VO distribution of Old English between 1150 A.D. and 1200 A.D.? To accomplish this, a continuous stream of Scandinavian speakers would be the force that caused the overall composition of the Old English population to drift towards a VO-biased distribution by 1150 A.D. These Scandinavians would learn Old English as a second language, and therefore likely learn it imperfectly, perhaps introducing a continuous VO bias into the data set available to learners in the population.

Old English learners, not filtering the input, would simply converge on exactly the distribution they encountered in the input from the mixture of native Old English and Scandinavian speakers using Old English as a second-language. This scenario, however, would require an exponential increase of incoming Scandinavians in order to get the gradual population-level shift before 1150 A.D. and the sharp population-level shift after 1150 A.D. This seems to be a rather unlikely event.

Still, there is another variant on this external factor. Suppose there was some prestige associated with the Scandinavians such that Old English speakers altered their OV/VO usage to accommodate (see Giles & Powesland (1975) for accommodation theory) and sound more like the Scandinavian portion of the population. So, Scandinavians would be learning Old English as a second language from native Old English speakers who would be more VO-biased (as a conscious social effort). The overall composition of the population would then be increasingly more VO-biased as time went on. Yet, in order to achieve the historical S-shaped trajectory of change, again there needs to be an exponential increase somewhere – either in the number of Scandinavians joining the Old English population or in the associated prestige with the Scandinavian VO-bias. While less unlikely than the previous scenario, relying on an exponential increase of Scandinavian prestige doesn't seem ideal as the sole factor driving change, either.

Nonetheless, we should not discount Scandinavian influence completely. Scandinavian influence combined with input filtering could well give the desired change. Later in this chapter, we will see that adult utterances generated with OV order are more prone than their VO counterparts to becoming ambiguous in the observable data. Scandinavian influence, being VO-biased, could have been responsible for this. Thus, learners using an unambiguous data filter would have become more VO-biased over time since the VO data generated by the Old English speakers was less likely to become ambiguous. Still, it is crucial to note that this scenario is the result of the combination of Scandinavian influence and language change caused by language learning. Scandinavian influence alone seems unlikely to be the cause of the language change in Old English.

4.3.3.2.2 Norman Influence

A second external source of influence is the Norman invasion in 1066 A.D. The Norman invaders spoke Old French, which was OV-biased in its distribution

(Kibler 1984): embedded clauses were predominantly OV order, as well as the matrix clauses. So, contact with Old French speakers would have biased the Old English population to become more OV. However, between 1000 and 1150 A.D., the Old English population was already drifting towards being more VO in its distribution. So, any contact with Old French speakers would have hindered the population-level change to a VO-biased distribution. This influence may have been tempered (and overcome) by the VO-biased Scandinavian influence.

Another possibility is disaccommodation with the OV-biased distribution from the Old French speakers if there was social stigma associated with the language of the Norman invaders (again, see Giles & Powesland (1975) for accommodation theory). Old English speakers, disliking the invaders (and perhaps liking the Scandinavians) would be driven to more VO-biased usage. Still, it remains clear that contact with the Normans alone could not have *caused* the shift in Old English to a strongly VO-biased distribution unless, as discussed for the Scandinavian influence in the previous section, there was an exponential increase somewhere – in this case, in the social stigma associated with using an OV-biased distribution.

4.4 The Model

I now describe the model at the individual level and the population level. Because I have posited that language change at the population level is driven by language learning at the individual level, I first examine the details of individual learning. In the model, the learner has different hypotheses for a structure in the language (such as OV and VO word order) available during learning, in line with work by Yang (2002), Dresher (1999), Lightfoot (1999), Fodor (1998a, 1998b), Niyogi & Berwick (1997, 1996, 1995), and Clark & Roberts (1993). The target state after learning is complete is a probabilistic distribution between competing hypotheses (Yang, 2002; Pintzuk, 2002; Kroch & Taylor, 1997; Bock & Kroch, 1989). Because of this, individual linguistic behavior, whether child (Yang, 2003) or adult (Bock & Kroch, 1989), is represented as a probabilistic distribution of multiple structural hypotheses, specifically between OV and VO word order.

Population-level change in the model is the result of a build-up of individual-level "mislearning" (Yang, 2002, 2000; Briscoe, 2000, 1999; Niyogi & Berwick, 1997, 1996, 1995; Clark & Roberts, 1993; Lightfoot, 1991). Thus, the population-level model relies heavily upon the individual-level implementation.

4.4.1 The Individual-Level Model

4.4.1.1 Learning in the Individual

The individual-level model is a model of language learning. An individual learner in the model is instantiated with a probability p_{VO} of accessing VO word order. The OV word order is accessed with probability $1 – p_{VO}$, as there are only two hypotheses under consideration.

In a language system where the adult speakers have $p_{VO} = 1.0$ (modern English) or $p_{VO} = 0.0$ (modern Dutch and German), all utterances are produced with one word order (VO for modern English, OV for modern Dutch and German). This directly impacts the distribution of unambiguous data, since all unambiguous data will be unambiguous for a single hypothesis (either OV words order or VO word order).

In contrast, a language system can also exist where the adult p_{VO} is greater than 0.0 and less than 1.0, such as the state of Old English between 1000 A.D. and 1200 A.D. In a system like Old English, VO order is accessed for production with probability p_{VO} (which is less than 1.0) and the OV order is accessed with probability 1-p_{VO} (which is greater than 0.0). This will impact the distribution of unambiguous data: the data will have some distribution between p_{VO} = 0.0 (all OV order data) and p_{VO} = 1.0 (all VO order data). The learner then determines her own p_{VO} based on the distribution in the intake (which, in the model, will be filtered down to the degree-0 unambiguous data).

The model assumes no initial bias for either hypothesis, so the initial value for a learner's word order, p_{VO}, is 0.5. This can be interpreted as an unbiased value, since it is precisely in the middle of p_{VO} = 0.0 (all OV order) and p_{VO} = 1.0 (all VO order). Note that an unbiased p_{vo} would predict that very young children of any language would have an unstable word order initially. I speculate that the reason why children always demonstrate knowledge of the correct word order by the time they reach the two word stage is because they have already been exposed to enough examples of the appropriate word order for their language to bias them in the correct way.

The final p_{VO} value after the learning period is complete will range between 0.0 and 1.0, and can be interpreted as a probabilistic access of the OV and VO words orders. A p_{VO} of 0.3, for example, would correspond to accessing VO order 30% of the time during production and OV order 70% of the time.

Since the initial p_{VO} for the learner is 0.5, the learner initially expects the distribution of OV and VO data in the intake to be unbiased. I use the Bayesian framework laid out in chapter 2 to model how the learner's initial hypothesis about the OV/VO distribution (p_{VO}) shifts with each additional data point from the intake. In addition to the support for its psychological validity in human cognition (Tenenbaum & Griffiths, 2001), Bayesian learning has also been used in other models of language evolution and change (Briscoe, 1999).

Since there are only two values for the OV/VO ordering (OV and VO), I represent the learner's hypothesis of the expected distribution of OV and VO utterances as a binomial distribution centered around some probability *p*. Here, probability *p* is p_{VO} and represents the learner's belief about the likelihood of encountering a VO utterance. When p_{VO} is 0.5, the learner is most confident that it is equally likely that an OV or the VO utterance will be encountered. A p_{VO} near 0.0 means the learner is most confident that a VO utterance will never be encountered; a p_{VO} near 1.0 means the learner is most confident that a VO utterance will always be encountered.

The learner's p_{VO} is updated by calculating the maximum of the a posteriori (MAP) probability of the prior belief p_{VOprev}, given the current piece of data from the intake. In essence, the model is starting with a prior probability and its expected distribution of OV and VO utterances, and comparing this expected distribution against the actual distribution encountered. The updated probability is calculated as follows:

(16a) If the data point is analyzed as OV, $p_{VO} = (p_{VOprev} * t)/(t+c)$
(16b) If the data point is analyzed as VO, $p_{VO} = (p_{VOprev} * t+c)/(t+c)$

where t = total expected number of data points in the *intake* during the period of fluctuation (2000 in this model) and c = learner's confidence in the input (ranging between 0.0 and 5.0), based on p_{VOprev}. Note that t refers to quantity of data points in the intake, and not the input. Thus, the learner will encounter considerably more than 2000 data points in the input; the fluctuation period, however, ends when 2000 data points from the intake have been encountered.

Also note that these equations are a modification of the update equations derived in chapter 2. In those equations, $c = 1$. However, I have modified this value since those equations (with $c = 1$) would not allow the learner to converge to 1.0 or 0.0, even if all unambiguous data are of one value. For example, with $t = 2000$, encountering all OV data points causes the final p_{VO} to be 0.194 (not 0.0); encountering all VO data points causes the final p_{VO} to be 0.816 (not 1.0). I therefore modified c to allow the final p_{VO} to be closer to the endpoint values (either 0.0 or 1.0) for each case.

The value c ranges linearly between 0 and a maximum value m, depending on what p_{VOprev} is[27]:

(17a) VO data: $c = p_{VOprev} * m$
(17b) OV data: $c = (1 - p_{VOprev}) * m$

The value m ranges between 3.0 and 5.0. The m for a particular mixture of degree-0 and degree-1 data is determined by seeing which m value allows the simulated Old English population to reach an average p_{VO} value in the population between 1000 and 1150 A.D that accords with the historical data available. For example, the value of m for an intake that consists only of degree-0 data is 5.0.

With the new update functions, unambiguous data for one value the entire time will cause the final p_{VO} to be much closer to the endpoint. Seeing 2000 OV data points leaves p_{VO} between .007 and .048 (depending on m); seeing 2000 VO data points leaves p_{VO} between .952 and .993 (depending on m).

The final p_{VO} at the end of the fluctuation period (after t data points from the intake have been encountered) will reflect the distribution of the data points in the intake. Importantly, the distribution is reflected *without* the learner explicitly memorizing each individual piece of data for later analysis. Instead, as each data point is encountered, the information is extracted from that data point and, using the equations in (16) and (17), integrated into the learner's hypothesis about what the distribution of OV and VO data points is expected to be.

The individual learning algorithm used in the model is described in (18):

[27] The same effect could likely be achieved by holding c between 0 and 1, and letting t vary. However, this loses the intuition that t (the number of data points the learner expects, i.e. the amount of change allowed) should be the same across the different conditions investigated.

(18) Individual learning algorithm

(a) Set initial p_{VO} to 0.5.

(b) Get a data point from an "average" member of the population. The input for the learner is determined by sampling from a normal distribution around the average p_{VO} of the population.

(c) If the data point is degree-0 and unambiguous, use this data point as intake and then alter p_{VO} accordingly.

(d) Repeat (b-c) until the fluctuation period is over, as determined by *t*.

For each data point encountered from the input, the learner determines if the data point belongs in the intake. If so, p_{VO} is updated using the equations in (16-17). This process of encountering input and integrating the information from data in the intake continues until the fluctuation period is over. At that point, the learner becomes one of the population members that contribute to the average p_{VO} value that will influence future learners. The higher the average p_{VO} value is in the population, the more likely learners are to encounter unambiguous VO data.

4.4.1.2 Old English Intake Data

As we have seen, the distribution in the learner's intake controls the learner's shift away from the unbiased probability of $p_{VO} = 0.5$. The only way to shift p_{VO} away from 0.5 is to have more data points of one word order than of the other in the intake. I will refer to this quantity as the bias one word order has over the other. [28] So, if the intake distribution is OV-biased, there are more OV data points in the learner's intake. If the intake distribution is VO-biased, there are more VO data points in the learner's intake. Note that if the intake is a subset of the input (due to filtering), the bias with respect to the available input is smaller than the bias with respect to the learner's intake. Table 4.1 displays the OV bias with respect to the input in the degree-0 (D0) and degree-1 (D1) clauses in Old English at various points in time.

	D0 Total # Clauses	D0 Unamb OV	D0 Unamb VO	D0 OV Bias w.r.t. the input$_a$	D0 OV Bias w.r.t. the intake$_b$
1000 A.D.	9805	1389	936	4.6%	19.5%

[28] This differs from the *advantage* (Yang, 2000) one hypothesis has over another. Advantage there is defined as inherent grammar incompatibility – one hypothesis will have an advantage when the opposing hypothesis is incompatible with data *types*. Thus, it does not matter for advantage how frequent a data type is, e.g. how many data *tokens* appear in the intake. It simply matters that there are data types one hypothesis is incompatible with. Advantage is thus different from the *bias* in the intake distribution, which very much depends on the quantity of data tokens that are unambiguous for one hypothesis vs. the other. More specifically, a hypothesis with a lower advantage can still have a stronger bias in the data intake distribution, and vice versa.

1000 – 1150 A.D	6214	624	590	0.5%	2.8%
1200 A.D.	1282	180	190	-0.8% [c]	-2.7% [c]

	D1 Total # Clauses	D1 Unamb OV	D1 Unamb VO	D1 OV Bias w.r.t. the input[a]	D1 OV Bias w.r.t. the intake[b]
1000 A.D.	7559	3844	1583	29.9%	41.7%
1000 – 1150 A.D	3636	1759	975	21.6%	28.7%
1200 A.D.	2236	551	1460	-40.7% [c]	-45.2% [c]

Table 4.1. OV order bias in the input for degree-0 (D0) and degree-1 (D1) clauses.
[a] The bias for the OV order with respect to the *input* is derived by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of clauses in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (1389-936)/9805 = 4.6%. [b] The bias for the OV order with respect to the *intake* is derived by subtracting the quantity of unambiguous VO data from the quantity of unambiguous OV data, and then dividing by the total number of clauses in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (1389-936)/(1389+936) = 19.5%. [c] Note that a negative OV bias means that the distribution is VO-biased.

The corpus data show a 4.6% bias with respect to the input for the OV order in the degreee-0 clauses at 1000 A.D. We can interpret this as less than 5 out of every 100 sentences of the available input are biasing the learner away from a p_{VO} of 0.5 (and towards an OV value of 0.0). With respect to the intake, the OV order bias is much higher: just about 1 out of every 5 data points in the intake biases the learner towards 0.0 (OV order).

Interestingly, the OV bias in the degree-1 clauses is much higher (29.9% with respect to the input, and 41.7% with respect to the intake). However, a degree-0 filter would cause the learner to ignore these data that would shift p_{VO} towards 0.0 significantly more often. Nonetheless, the difference of the bias in the different distributions highlights the effect that data intake filtering can have: the bias in the distribution alters quite a lot depending on which data set the learner is using.

4.4.2 Population-Level Model for Old English

4.4.2.1 Population-Level Algorithm and Population Growth

The population algorithm (19) centers on the individual acquisition algorithm in (18).

(19) Population-level algorithm

(a) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(b) Initialize the members of the population to the average p_{VO} at 1000 A.D.
(c) Set the time to 1000 A.D.
(d) Move forward 2 years.
(e) Members age 59-60 die off. The rest of the population ages 2 years.
(f) New members are born. These new members use the individual acquisition algorithm (18) to set their p_{VO}.
(g) Repeat steps (d-f) until the year 1200 A.D.

The population members range in age from newborn to 60 years old.[29] The initial size of the population is 18,000, based on estimates from Koenigsberger & Briggs (1987). At 1000 A.D., all the members of the population have their p_{VO} set to the same initial p_{VO}, which is derived from the historical corpus data. Every two years, new members are born to replace the members that died as well as to increase the overall size of the population so it matches the growth rate extrapolated from Koenigsberger & Briggs (1987). Populations are estimated to grow at an exponential rate characterized by the equation in (20).

(20) Population growth equation

$$\text{population size} = \text{previous population size} * e^{rt}$$

For the Old English population in our model, $r = 0.00400953$ and t = time in years. For example, at 1002 A.D., the estimated population size is $18000*e^{0.00400953*2} = 18145$. Thus, once the oldest members (age 59-60) die off, enough new members are born to make the total population size at 1002 A.D. be 18145. These new members encounter input from the rest of the population and follow the process of individual acquisition laid out previously in order to determine their final p_{VO}. This process of death, birth, and learning continues until the year 1200 A.D.

4.4.2.2 Population Values from Historical Data

I use the historical corpus data to initialize the average p_{VO} in the population at 1000 A.D., calibrate the model between 1000 and 1150 A.D. (recall that the confidence value c in update equation (16) needs calibration), and determine how strongly VO-biased the distribution has to be in the population by 1200 A.D. But it is not straightforward to determine the average p_{VO} at a given period of time.

Both the degree-0 and degree-1 unambiguous data distributions are likely to be distorted from the underlying unambiguous data distribution produced by p_{VO} because the degree-0 and degree-1 clauses have ambiguous data. The underlying

[29] The population members begin uniformly distributed between 0 and 60 years old, though this could easily be modified to a more skewed distribution where there are more younger members of the population than older. In addition, the age maximum (60 years old) was arbitrarily chosen. Having a lower maximum (say, 40 years old) would possibly speed the rate of change through the population. However, the overall results would likely be the same as found here since the population model must be calibrated so that the population remains sufficiently OV-biased before 1150 A.D. That is, a sufficient OV-bias in the population before 1150 A.D. is a precondition. The behavior we are interested in is how a population that is sufficiently OV-biased before 1150 A.D. changes between 1150 A.D. and 1200 A.D. Specifically, can it become VO-biased enough?

distribution in a speaker's mind, however, has no ambiguous data – every clause is generated with OV or VO order. As we can see in table 4.2, the degree-0 clauses have more ambiguous data than the degree-1 clauses. Moreover, recall from table 1 that the degree-1 clauses also have a magnified bias, compared to the degree-0 clauses. Taken together, I use these two observations to make the assumption that the degree-0 distribution is more distorted than the degree-1 distribution.

	D0 Total # Clauses	D0 # Unamb Clauses	D0 % Ambig[a]
1000 A.D.	9805	2325	(9805-2325)/9805 = 76%
1000-1150 A.D.	6214	1214	(6214-1214)/6214 = 80%
1200 A.D.	1282	370	(1282-370)/1282 = 71%

	D1 Total # Clauses	D1 # Unamb Clauses	D1 % Ambig[a]
1000 A.D.	7559	5427	(7759-5427)/7759 = 28%
1000-1150 A.D.	3636	2734	(3636-2734)/3636 = 25%
1200 A.D.	2236	2011	(2236-2011)/2236 = 10%

Table 4.2. Percentage of ambiguous clauses in the historical corpora. [a] The % Ambig is calculated by dividing the number of ambiguous clauses (Total - Unamb) by the total number of clauses.

I then use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the degree-1 distribution and the underlying unambiguous data distribution in a speaker's mind. In this way, I estimate the underlying unambiguous data distribution (produced by p_{VO}) for an average Old English speaker at certain points in time.

I will first step through the formalization of the procedure used to derive the underlying p_{VO} at a given point in time. Then, I will step through an explicit example from the Old English historical data.

4.4.2.2.1 Procedure to Derive p_{VO} from Historical Data

Let there be two hypotheses under consideration, h1 and h2. For Old English, these are OV order (h1) and VO order (h2). From historical corpora, we can gather unambiguous data points for h1 and h2 in both the degree-0 and degree-1 clauses. From these, we can calculate the number of ambiguous data points in the degree-0 and degree-1 clauses. The quantities gathered from historical corpora are **u1d0** (**u**nambiguous data points for h**1** in **d**egree-**0** clauses), **u2d0** (**u**nambiguous data points for h**2** in **d**egree-**0** clauses), **ad0** (**a**mbiguous data points in **d**egree-**0** clauses), **u1d1** (**u**nambiguous data points for h**1** in **d**egree-**1** clauses), **u2d1** (**u**nambiguous data points for h**2** in **d**egree-**1** clauses), and **ad1** (**a**mbiguous data points in **d**egree-**1** clauses) in table 4.3 below. The quantities that must be derived are *u1* and *u2*, which represent the quantities of unambiguous data for each hypothesis in the underlying distribution that the average population speaker produced. In the underlying distribution, there are no ambiguous data because the speaker either accesses h1 or h2 to produce the data point. Once *u1* and *u2* are known, p_{VO} can be derived ($p_{VO} = u2/(u1 + u2)$).

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1**	**u2d1**	**ad1**
Underlying Distribution	*u1*	*u2*	0

Table 4.3. Formalization of quantities available from historical corpora and quantities to derive. Quantities in **bold** can be gathered from historical corpora. Quantities in *italics* must be derived and are used to calculate the average p_{VO} in the population.

Let γ represent the probability that the speaker accesses h1 during production. Since there are only two options under consideration, 1 - γ represents the probability the speaker accesses h2 during production.

Let the total quantity of degree-0 data be **d0**. So, **d0** = **u1d0 + u2d0 + ad0**.

Let the total quantity of degree-1 data be **d1**. So, **d1** = **u1d1 + u2d1 + ad1**.

We first must normalize the degree-1 data quantity to the degree-0 data quantity. After normalization, **u1d1' + u2d1' + ad1' = d0 = u1d0 + u2d0 + ad0**.

(21) Equation quantities, original and normalized

(a) **d0** = **u1d0 + u2d0 + ad0**

(b) **d1** = **u1d1 + u2d1 + ad1**

(c) **d0 = u1d1' + u2d1' + ad1'**

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1'** **= u1d1*(d0/d1)**	**u2d1'** **= u2d1*(d0/d1)**	**ad1'** **= ad1*(d0/d1)**
Underlying Distribution	*u1*	*u2*	0

Table 4.4. Data quantities after normalization.

The value *u1* represents the quantity of unambiguous h1 (OV) data generated by the speaker. The value *u2* represents the quantity of unambiguous h2 (VO) data generated by the speaker. Since there are no ambiguous data, let these two quantities also sum to **d0** (*u1* + *u2* = **d0**). This represents the intuition that *u1* and *u2* have been "normalized" so that they can be compared against their counterpart values in the degree-1 and degree-0 distributions. Note that since *u1* and *u2* have not been calculated yet, we can simply make them sum to the appropriate normalized value, **d0**.

(22) Underlying distribution "normalization"

u1 + *u2* = **d0**

Recall that the probability that a speaker accesses h1 when producing a data point is γ. Since the total quantity of unambiguous data points in the underlying distribution has been normalized to **d0**, this probability can now be set equal to *u1*/**d0**. Thus, we can rewrite *u1* as γ***d0**.

(23) Rewriting underlying distribution quantity *u1*

$\gamma = u1/\mathbf{d0}$

$u1 = \gamma*\mathbf{d0}$

I now make an assumption about the relation of underlying data distribution to the degree-1 data distribution. Specifically, I assume that the degree-1 distribution originally had the same number of h1 data points as the underlying distribution, but that some of these data points became ambiguous (due to various grammatical operations). Thus, we can relate the underlying distribution h1 data point quantity *u1* to the degree-1 data quantities **u1d1’** (normalized quantity of **u**nambiguous data points for h**1** in the **d**egree-**1** distribution) and **ad1’** (normalized quantity of **a**mbiguous data points in the **d**egree-**1** distribution).

(24) Relation between *u1* and **u1d1’** and **ad1’**

u1 = **u1d1’** + the portion of **ad1’** that were originally h1 data points

Let *a1d1* = portion of **ad1’** that were originally h1 data points

u1 = **u1d1’ +** *a1d1*

Recall from (23) that *u1* can be rewritten in terms of γ and **d0**. We can thus write an equation for *a1d1*, the portion of **ad1’** that were originally h1 data points.

(25) Writing an equation for *a1d1*

$u1 = \gamma*\mathbf{d0}$ (from (23))

$\gamma*\mathbf{d0} = \mathbf{u1d1'} + a1d1$ (from (24))

$a1d1 = \gamma*\mathbf{d0 - u1d1'}$

Since there are only two hypotheses, the portion of **ad1’** that were not originally h1 data points must have been h2 data points. Given this, we can write an equation for *a2d1*, the portion of **ad1’** that were originally h2 data points.

(26) Writing an equation for *a2d1*

Let *a2d1* = portion of **ad1’** that were originally h2 data points

$\mathbf{ad1'} = a1d1 + a2d1$

$a2d1 = \mathbf{ad1'} - a1d1$

Moreover, using the same assumption as before about the relation between the underlying distribution and the degree-1 distribution, we can rewrite *u2*, the quantity of unambiguous data points for h2 in the underlying distribution.

(27) Rewriting *u2*

u2 = **u2d1’** + the portion of **ad1’** that were originally h2 data points

a2d1 = portion of **ad1’** that were originally h2 data points

u2 = **u2d1’ +** *a2d1*

$u2 = \mathbf{u2d1'} + \mathbf{ad1'} - a1d1$ (from (26))

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**u1d0**	**u2d0**	**ad0**
Degree-1	**u1d1'** **= u1d1*(d0/d1)**	**u2d1'** **= u2d1*(d0/d1)**	**ad1' =** **ad1*(d0/d1)**
Underlying Distribution	**u1d1'** + *a1d1*	**u2d1'** + (**ad1'** – *a1d1*)	0

Table 4.5**.** Derived quantities rewritten.

Now, we look at the relation between the degree-1 and the degree-0 distribution. I make an assumption similar to the one we did about the relation between the underlying distribution and the degree-1 distribution: specifically, I assume that the degree-0 distribution originally had the same number of h1 or h2 data points as the degree-1 distribution, but that some of these data points became ambiguous (due to various grammatical operations). I can describe these quantities in terms of values we have already observed or calculated.

I assume that the quantity of h1 data points in the degree-0 distribution was originally the same as the quantity of h1 data points in the normalized degree-0 distribution, **u1d1'**. However, some became ambiguous and only **u1d0** remain. So, the quantity of data points that became ambiguous going from the degree-1 distribution to the degree-0 distribution can be described as **u1d1'** – **u1d0**. The same reasoning can be used for the h2 data points.

(28) Quantities of data points that became ambiguous going from the degree-1 distribution to the degree-0 distribution

Let the quantity of h1 data points that became ambiguous going from the degree-1 to the degree-0 distribution = *a1d1to0.*

$a1d1to0 = \mathbf{u1d1'} - \mathbf{u1d0}$

Let the quantity of h2 data points that became ambiguous going from the degree-1 to the degree-0 distribution = *a2d1to0*

$a2d1to0 = \mathbf{u2d1'} - \mathbf{u2d0}$

We can now define an *ambiguity loss ratio* **Ld1to0**, which represents the ratio of h1 data points that became ambiguous compared to the h2 data points that became unambiguous going from the degree-1 to the degree-0 distribution.

(29) Ambiguity Loss Ratio **Ld1to0**
(h1 data point loss over h2 data point loss going from degree-1 to degree-0 distribution)

$$\mathbf{Ld1to0} = \frac{\mathbf{u1d1' - u1d0}}{\mathbf{u2d1' - u2d0}}$$

We can then describe the quantities of h1 and h2 data points that become ambiguous going from the underlying distribution to the degree-1 distribution. Let *a1utod1* be the quantity of h1 data points that became ambiguous going from the underlying distribution to the degree-1 distribution. Let *a2utod1* be the quantity of h2 data points that become ambiguous going from the underlying distribution to the degree-1 distribution.

(30) Describing the quantities of h1 and h2 data points that become ambiguous going from the underlying to the degree-1 distribution

(a) *a1utod1* (h1 data points that become ambiguous)

$a1utod1 = u1 - \mathbf{u1d1'}$

$a1utod1 = (\mathbf{u1d1'} + a1d1) - \mathbf{u1d1'}$ (from (24))

$a1utod1 = a1d1$

(b) *a2utod1* (h2 data points that become ambiguous)

$a2utod1 = u2 - \mathbf{u2d1'}$

$a2utod1 = (\mathbf{u2d1'} + (\mathbf{ad1'} - a1d1)) - \mathbf{u2d1'}$ (from (27))

$a2utod1 = \mathbf{ad1'} - a1d1$

We can now define an ambiguity loss ratio **Lutod1**, which represents the ratio of h1 to h2 data points that become "lost" to ambiguity going from the underlying distribution to the degree-1 distribution. I make an assumption that **Ld1to0** is the same as **Lutod1**, that is that the rate at which h1 data points become ambiguous compared to h2 data points does not change depending on which distributions are being compared. For example, if h1 data points are twice as likely as h2 data points to become ambiguous going from the degree-1 to the degree-0 distribution, then I assume h1 data points are twice as likely as h2 data points to become ambiguous going from the underlying distribution to the degree-1 distribution.

(31) Ambiguity Loss Ratio Assumption

$$\mathbf{Lutod1} = \mathbf{Ld1tod0} = \frac{\mathbf{u1d1' - u1d0}}{\mathbf{u2d1' - u2d0}}$$

Now, we have all the pieces in place to write an equation that relates the ambiguity loss of h1 data points to the ambiguity loss of h2 data points going from the underlying distribution to the degree-1 distribution. The intuition is laid out in (32).

(32) Intuition to relate ambiguity loss from underlying to degree-1 distribution

% of h1 data points "lost" = **Lutod1*** % of h2 data points "lost"

$$\frac{\text{\# of h1 data points lost}}{\text{total \# of h1 data points}} = \mathbf{Lutod1} * \frac{\text{\# of h2 data points lost}}{\text{total \# of h2 data points}}$$

This intuition can be instantiated as in (33). We can then use the equations we have already derived to solve for γ, the probability of accessing h1 in the underlying distribution.

(33) Solving for γ

(from (32)) $\frac{a1utod1}{u1} = \mathbf{Lutod1} * \frac{a2utod1}{u2}$

(from (31)) $\frac{a1utod1}{u1} = \mathbf{Ld1tod0} * \frac{a2utod1}{u2}$

(from (25)) $\frac{a1utod1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{a2utod1}{u2}$

(from (27)) $\frac{a1utod1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{a2utod1}{\mathbf{u2d1'} + \mathbf{ad1'} - a1d1}$

(from (30)) $\frac{a1d1}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{\mathbf{ad1'} - a1d1}{\mathbf{u2d1'} + \mathbf{ad1'} - a1d1}$

(from (25)) $\frac{\gamma * \mathbf{d0} - \mathbf{u1d1'}}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{\mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}{\mathbf{u2d1'} + \mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}$

$$\begin{aligned}&\gamma^2(\mathbf{Ld1tod0} + 1)(\mathbf{d0^2})\\ &\quad + \gamma(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))\\ &\quad + (\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}) = 0\end{aligned}$$

Now, we can use the quadratic formula to solve for γ.

$$\begin{aligned}a &= (\mathbf{Ld1tod0} + 1)(\mathbf{d0^2})\\ b &= (\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))\\ \mathrm{c} &= (\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'})\end{aligned}$$

$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0^2})}$$
$$+/- \frac{\sqrt{((\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'})))^2 - 4(\mathbf{Ld1tod0} + 1)(\mathbf{d0^2})((\mathbf{-1})(\mathbf{d0} * \mathbf{u1d1'}))}}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0^2})}$$

This formula can be easily resolved once we insert the observable quantities from the historical corpus, as we will see in the next section. Once we have solved for γ, we have the probability with which h1 is accessed in the underlying distribution. We can calculate the probability with which h2 is accessed in the underlying distribution by using 1 - γ.

4.4.2.2.2 A Concrete Example of Deriving p_{VO} from Historical Data

For our Old English corpus, let h1 be the OV word order hypothesis and h2 be the VO word order hypothesis. I will step through the derivation of the underlying p_{VO} value at 1000 A.D. First, we observe the various quantities available from the historical corpus at 1000 A.D.

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**1389**	**936**	**7480**
Degree-1	**3844**	**1583**	**2132**
Underlying Distribution	*u1*	*u2*	0

Table 4.6. Quantities available from historical corpora and quantities to derive. Quantities in **bold** are gathered from historical corpora. Quantities in *italics* must be derived and are used to calculate the average p_{VO} in the population.

Then, we normalize the degree-1 quantities to the degree-0 quantities. The total quantity of degree-0 data **d0** is 1389 + 936 + 7480 = **9805**. The total quantity of degree-1 data **d1** is 3844 + 1583 + 2132 = **7559**. To normalize the degree-1 quantities, we therefore multiply each quantity by **(d0/d1)** = (**9805/7559**).

	Unamb OV (h1)	Unamb VO (h2)	Amb
Degree-0	**1389**	**936**	**7480**
Degree-1	**4986**	**2053**	**2766**
Underlying Distribution	*u1*	*u2*	0

Table 4.7. Data quantities after normalization.

We then calculate the ambiguity loss ratio between the degree-1 and degree-0 distribution, **Ld1tod0**.

$$(34)\ \mathbf{Ld1tod0} = \frac{\mathbf{u1d1' - u1d0}}{\mathbf{u2d1' - u2d0}} = \frac{\mathbf{4986 - 1389}}{\mathbf{2053 - 936}} \approx 3.22$$

So, we see that the OV data points are over three times as likely to become ambiguous as the VO data points at 1000 A.D. I assume that this loss ratio is the same going from the underlying distribution to the degree-1 distribution (**Lutod1**), that is that OV data points are three times as likely as VO data points to become ambiguous going from the underlying to the degree-1 distribution.

We now have all the quantities we need to calculate γ (from (33)).

$$\gamma = \frac{\textbf{-(9805)(9805 + 4986 - 3.22*(2766 + 4986))}}{2(\textbf{3.22} \ +1)(\textbf{9805}^2)}$$

$$+/- \frac{\sqrt{((\textbf{9805})(\textbf{9805 + 4986 - 3.22*(2766 + 4986)}))^2 - 4(\textbf{3.22} \ +1)(\textbf{9805}^2)((\textbf{-1})(\textbf{d0*4986'}))}}{2(\textbf{3.22} \ +1)(\textbf{9805}^2)}$$

Solving for γ, we obtain 0.766 and -.299. Since we know γ is a probability and must be between 0.0 and 1.0, the correct solution for γ is 0.766. So, given these historical data distributions, I estimate that the OV word order option was accessed with probability 0.766 at 1000 A.D. The VO word order option was thus accessed with probability 1-0.766 = 0.234. Since we are tracking the probability with which the VO word order option is accessed, p_{VO} is 0.234 at 1000 A.D. The average p_{VO} values in the population at the other two periods of time we consider (1000-1150 A.D. and 1200 A.D.) can be calculated in the same fashion by using the quantities in table 4.1. Table 4.8 below displays the three p_{VO} values I will be using in my simulations.

	Degree-0 Clauses			Degree-1 Clauses			Underlying
	Total	OV Unamb	VO Unamb	Total	OV Unamb	VO Unamb	p_{VO}
1000 A.D.	9805	1389	936	7559	3844	1583	**.234**
1000 – 1150 A.D.	6214	624	590	3636	1759	975	**.310**
1200 A.D.	1282	180	190	2236	551	1460	**.747**

Table 4.8. Data from historical corpora and calculated p_{VO}.

To model the data from the historical corpus, a population must start with an average p_{VO} of 0.234 at 1000 A.D., reach an average p_{VO} of 0.310 between 1000 and 1150 A.D.[30], and reach an average p_{VO} of 0.747 by 1200 A.D.

4.4.2.3 Answering Questions About Learning Filters

Armed with these models of individual-level learning and population-level change, we can now answer two questions about filters on the learner's intake. First, I address the question of descriptive *sufficiency*: can an Old English population whose learners filter their intake down to the degree-0 unambiguous data shift from a strongly OV biased distribution to a strongly VO biased distribution at the appropriate time? Recall that the data intake set is significantly smaller than the data input set, and so there is a potential data sparseness problem. Moreover, recall that exactly the right amount of misconvergence on the p_{VO} value must happen for each set of new population members in order for the population as a whole to change at the correct

[30] This is what is meant by calibration. If the population is unable to reach this checkpoint, it is unfair to compare its p_{VO} at 1200 A.D. against other populations' p_{VO} values at 1200 A.D. The value which must be calibrated is the learner's confidence value *c* in the current piece of data, which determines how much the current p_{VO} is updated for a given data point.

rate. We can ask if input filtering in the specified manner can cause this precise amount of misconvergence.

Second, we address the question of *necessity*. If the proposed intake filtering is sufficient to cause an Old English population to change at the correct rate, is it in fact necessary? One might wonder if an Old English population that does not use either filter, or only uses one (either unambiguous data or degree-0), would achieve the same results. With the model described here, we can find out.

4.5 Modeling Results

4.5.1 Sufficient Filtering

We first examine the descriptive sufficiency of the data intake filters. Does our simulated Old English population, whose learners filter their intake down to the degree-0 unambiguous data, undergo change at the historically attested rate? Figure 29 shows the average population p_{VO} over time. Based on these simulation data, an Old English population using these filters can indeed shift from a strongly OV-biased distribution to a strongly VO-biased distribution at the historically correct time. Specifically, these filters yield a data set with the right bias at each point in time. This then allows individual learners in the population to misconverge exactly the right amount, which then leads to population-level change at the correct rate.

Moreover, we can see that the concern over data sparseness can be put aside. Despite the significantly smaller quantity of data that comprises the intake for these learners, the trajectory of the population is still in line with the known historical trajectory. We also note that the S-shaped curve so often observed in language change (Bailey, 1973; Weinreich, Labov, & Herzog, 1968; Osgood & Sebeok, 1954; among others) emerges here from the learners filtering their input and the subsequent small changes spreading through an exponentially growing population.[31]

[31] As mentioned previously, this demonstrates that external factors are not *necessary* to cause swift population-level change. Here, the population-level change results from internal factors: the language-learning mechanism at the individual-learning level.

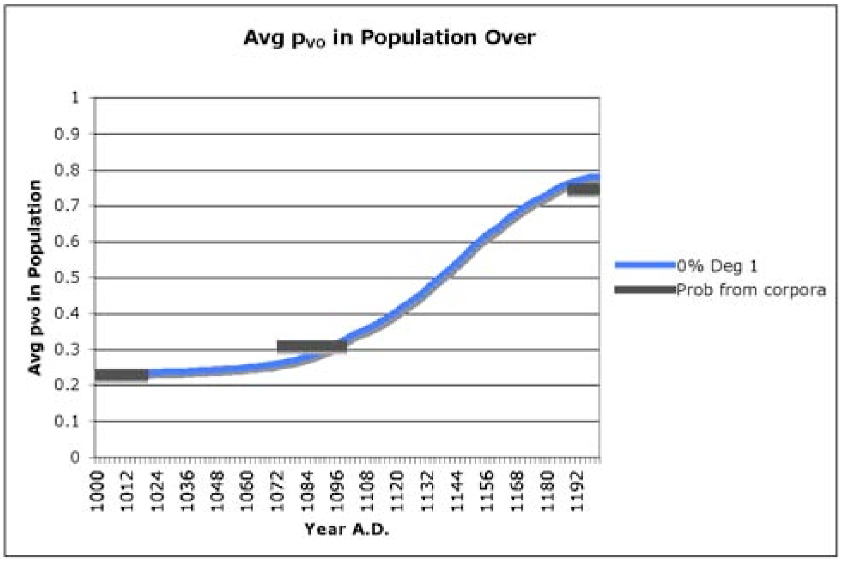


Figure 29. The trajectory of a population learning only from degree-0 unambiguous data, compared against estimates from historical corpora.

4.5.2 Necessary Filtering

We have just seen that these data intake filters are sufficient to cause the right rate of population-level change to occur. But are they necessary? Specifically, we wish to know if language change can occur at the historically attested rate without these filters. I examine the effects of removing each filter in turn, and then the effects of removing both.

4.5.2.1 Removing the Unambiguous Data Filter

I examine the unambiguous data filter first. A model could reasonably choose to drop this filter and assume that a learner attempts to activate the update algorithm for data that are ambiguous. In particular, the learner then requires some strategy to extract information from a given ambiguous data point. One simple strategy is for the learner to have a preference for analyzing strings as base-generated. This strategy would cause the learner to discard any analyses involving movement (for example, V2 movement) until forced to do so (Fodor, 1998b).

The effect of this strategy for the OV/VO word order cases we consider in Old English is that many more data points are used by the learner. Primary among these new data points are those of the form *Subject TensedVerb Object*. When V2 movement was considered in the analysis, this was ambiguous between OV order (OV, +V2) and VO order (VO, +/-V2), as we saw in example (13). However, if non-movement analyses are given preference, then the learner would take this ambiguous data point as evidence in favor of the VO word order hypothesis. Table 4.9 displays the data intake distribution for a learner who does not use an unambiguous data filter, as well as the OV word order bias at different points in time.

	D0 Total # Clauses	OV Data Intake	VO Data Intake	D0 OV Bias w.r.t. the input[a]	D0 OV Bias w.r.t. the intake[b]
1000 A.D.	9805	2537	3889	-13.8%	**-21.0%**[c]
1000 – 1150 A.D	6214	1221	2118	-14.4%	**-26.9%**
1200 A.D.	1282	389	606	-16.9%	**-21.8%**

Table 4.9. OV order bias in the degree-0 (D0) clauses. [a] We derive the bias for the OV order with respect to the *input* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/9805 = 13.8%. [b] We derive the bias for the OV order with respect to the *intake* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/(2537+3889) = 21.0%. [c] Note that a negative OV bias means that the distribution is VO-biased.

A very serious problem becomes apparent: even at the earliest time period when the population is supposed to be strongly OV-biased, the data intake distribution strongly favors the VO order. The VO word order has a 21.0% bias in the data intake at 1000 A.D (and a 13.8% bias in the input). Thus, about 21 out of every 100 data points encountered in the intake are biasing the learner towards the VO hypothesis. A population of learners using this data intake distribution could not remain strongly OV-biased for very long, and certainly not until 1150 A.D.

Therefore, I conclude that dropping the unambiguous data filter in this way will not allow the model to simulate what is actually observed in the Old English population. So, these results suggest that the unambiguous data filter is necessary.[32]

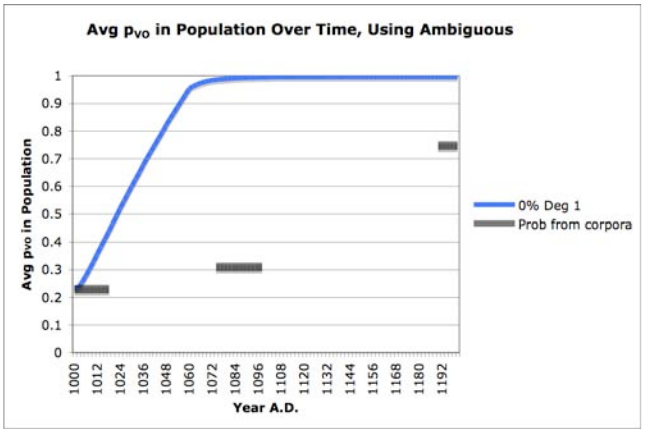


Figure 30. The trajectory of a population learning only from degree-0 data (ambiguous and unambiguous), compared against estimates from historical corpora.

[32] Unless we can find a strategy to deal with ambiguous data which includes a different set of data as intake, or values the ambiguous data in a manner that gives the OV hypothesis the advantage early on. The strategy explored here was the simplest (and most justifiable) one I could devise, but there may be more complex strategies that yield the desired results. If so, then we would need an explanation for the learner's knowledge and adoption of these more complex strategies.

4.5.2.2 Removing the Degree-0 Filter

I turn now to the degree-0 data filter. Suppose we drop this filter and allow the modeled learner to activate the update algorithm for both matrix (degree-0) and embedded (degree-1) clauses. Note that this learner still has the unambiguous data filter, and so will only activate the update procedure if the learner perceives the data point as unambiguous. Recall from table 4.1 that the degree-1 data intake distribution has a much higher OV bias before 1150 A.D. (28.7 – 41.7%). Given how high this OV bias is, it is possible that if there were enough degree-1 data in the input set, the learner would converge on a final p_{VO} that is too OV-biased. This slows the rate of change from OV-biased to VO-biased, and so a population made up of such learners would proceed much more slowly towards becoming VO-biased. I have estimated from the historical record that the Old English population should have an average p_{VO} value of 0.747 at 1200 A.D. This is the mark a simulated population must then reach.

With the model presented here, we can test the population-level effects of different compositions of data in the input set of the individual learner. Specifically, we can see how much (strongly OV-biased) degree-1 data can be in the input (and thus in this learner's intake) and still have the population as a whole be VO-biased enough by 1200 A.D. We can then compare this threshold against the estimated amount of degree-1 data available to learners and see if the degree-0 data filter is necessary. If the estimated amount of degree-1 data available to learners is less than the permissible threshold that allows correct population-level behavior, then the degree-0 filter is not necessary. The same population-level results can be obtained with or without the filter. In contrast, if the estimated amount of degree-1 data available to learners is greater than the permissible threshold, then we have support for the necessity of the degree-0 filter. This is because only by ignoring the degree-1 data available in the input can correct population-level behavior be obtained.

Figure 31 displays the average p_{VO} in the population at 1200 A.D. for 6 Old English populations whose learners had their input composed of different percentages of degree-1 data. For these populations, all the degree-1 data was in the intake set. Thus, a population with 16% degree-1 data in the input set activated the updating procedure for the 84% of the unambiguous data points that were degree-0 and the 16% of the unambiguous data points that were degree-1. Data points that were ambiguous were ignored.

The modeling results suggest that having even 4% degree-1 data available in the input (and thus in the learner's intake) is enough to prevent the simulated Old English population from reaching an average p_{VO} of 0.747 by 1200 A.D. We must now compare this threshold to the estimated amount of degree-1 data in the input to Old English learners.

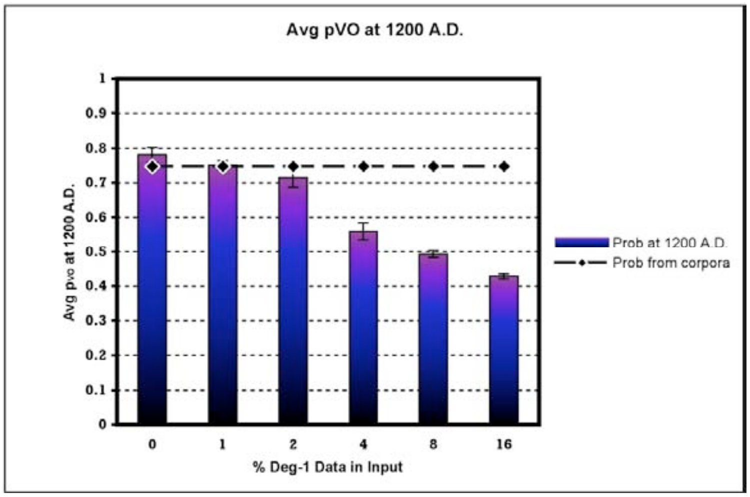


Figure 31. Average probability of using VO order at 1200 A.D. for populations with differing amounts of degree-1 data available during learning, as compared to the estimated average from historical corpora. Confidence intervals of 95% are shown.

I assume that amount of degree-1 child-directed data is approximately the same no matter what the time period (and I am currently unaware of studies that suggest otherwise). Given this, we can examine samples of modern English child-directed data to see what its composition is. The two samples I chose were a portion of the CHILDES database (MacWhinney, 2000) and some young children's stories (some of which can be found at http://www.magickeys.com/books/index.html). I used CHILDES since it is recorded speech to children and young children's stories because it is (storytelling) language designed to be directed at children. As we can see from Table 4.10, the CHILDES sample has approximately 8.8% degree-1 data points while the young children's stories sample has approximately 23.9% degree-1 data points. I take the average of these two sources to get an estimate of about 16% degree-1 data available in children's input. This is very similar to the 15% degree-1 data estimate from Sakas (2003), who examined several thousand sentences from the CHILDES database.

The modeling results (see figure 31) show that input comprised of 16% degree-1 data causes the simulated Old English population to be far too slow in shifting to a strongly VO-biased distribution. This is much higher than the permissible threshold of approximately 2%. Unless there is a way for the learner to allow in only an eighth of the degree-1 data available in the input, these results suggest that the degree-0 data intake filter is also necessary.[33]

[33] Another option is for the learner to weight the degree-1 data's influence so it is only an eighth as strong as the degree-0's influence. This particular weighting would then have to be justified.

A subsection of CHILDES				
Total Utterances	Total Data Points[a]	Total D0	Total D1	% D1
4068	2760	2516	244	8.8
Sample D0 Utterances		Sample D1 Utterances		
"What's that?", "I don't know.", "There's a table.", "Can you climb the ladder?", "Shall we stack these?", "That's right."		"I think it's time…", "Look what happened!", "I think there may be one missing.", "Show me how you play with that.", "See if you can get it.", "That's what he says."		

Young Children's Stories				
Total Utterances	Total Data Points[a]	Total D0	Total D1	% D1
4031	3778	2955	927	23.9
Sample D0 Utterances		Sample D1 Utterances		
"Ollie is an eel.", "She giggled.", "…but he climbs the tree!", "This box is too wide.", "…to gather their nectar."[b], "This is the number six."		"…that even though he wishes hard,…", "…that only special birds can do.", "…that can repeat words people say.", "…when the sun shines.", "…that goes NEIGH…NEIGH…", "…know what it is?"		

Table 4.10. Data gathered from speech directed to young children. [a] The number of data points is much less than the number of utterances since many of these utterances include "Huh?" and exclamations like "A ladder!" in the case of the spoken CHILDES corpus. For the young children's stories, there are often "sentences" like "Phew!" and "Red and yellow and green" which were excluded under Total Data Points. [b] I note that clauses with infinitives such as "…to gather their nectar" are included under degree-0 data, based on Lightfoot's (1991) definition of clause-union structures as degree-0. If this were not the case, the percentage of degree-1 clauses would only be higher than what I have calculated here – thus, this is a lower bound on the amount of degree-1 data available in the input.

4.5.2.3 Removing Both Filters

We have just observed that the loss of each of the data intake filters has a different effect on the rate of change at the population-level. Without the unambiguous data filter, the intake distribution is too heavily VO-biased. The population becomes strongly VO-biased too soon, and so changes too quickly. Without the degree-0 data filter, the intake distribution is too heavily OV-biased. The population becomes strongly VO-biased too late, and so changes too slowly. Given these opposite effects, one might wonder if dropping both filters would allow the simulated population to change at the correct rate. We must again examine the data intake distributions that learners would be using to see the effects of removing both filters.

	Total # Clauses	OV Data Intake	VO Data Intake	D0 OV Bias w.r.t. the input[a]	D0 OV Bias w.r.t. the intake[b]
Degree-0 Data	9805	2537	3889	**-13.8%**	**-21.0%**[c]
Degree-1 Data	7559	4650	2610	26.9%	28.1%

Table 4.11. OV order bias at 1000 A.D. with no filters. [a] We derive the bias for the OV order with respect to the *input* by subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the input. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/9805 = 13.8%. [b] We derive the bias for the OV order with respect to the *intake* by

subtracting the quantity of VO data from the quantity of OV data, and then dividing by the total number of data points in the intake. For instance, the D0 OV bias at 1000 A.D. is calculated as (2537-3889)/(2537+3889) = 21.0%. ^{c} Note that a negative OV bias means that the distribution is VO-biased.

In order for the Old English population to remain strongly OV-biased before 1150 A.D., the data intake distribution must at least be OV-biased at 1000 A.D. As we can see from table 4.11, the degree-0 data intake is heavily VO-biased (21.0% VO data bias). In order to drop the VO bias in the intake down to zero (so the OV order has at least a fighting chance with learners at 1000 A.D.), about 43% of the intake would need to consist of degree-1 data.

My estimate of the available amount of degree-1 data in child-directed data suggests that *less than half* of this amount of degree-1 data is available, at best (16%).[34] So, I conclude that we cannot drop both the unambiguous data filter and the degree-0 data filter, lest the population be driven to become strongly VO-biased too soon. The claim that both data intake filters are necessary is thus strengthened.

<u>*4.6 General Discussion*</u>

4.6.1 Necessary Filters

The results presented here serve as an existence proof that a population model whose individual learners employ data intake filtering can handle the specific case of word order change in Old English. The two critical filters are (a) use only data perceived as unambiguous and (b) use only degree-0 data. This means that the update procedure is only activated when data points obeying these constraints are encountered. Otherwise, the update procedure is not activated and the data points are effectively ignored for the purposes of learning.

I now examine what effects input filtering in general could have on language change, as well as the feasibility of input filtering.

4.6.2 Intake Filtering and Language Change

The nature of the input filter may be what differentiates situations of language change from situations of stable variation. If the intake becomes too mixed for the child to converge on the same probability weighting as the adult, then language change will occur. In cases where only one structural option is used in the adult population (as is often the case), the adult probability distribution will be 0.0 or 1.0. Given children's tendency to generalize to an extreme value from noisy data (Hudson Kam & Newport, 2005), the intake would have to be quite mixed in order to force children away from the adult distribution.

In this way, we see that learning can tolerate some variation in the input without causing the language to change. In this, our model's behavior differs notably

[34] Moreover, since not all the data in the input becomes intake, even more than 43% of the input would need to consist of degree-1 data. Give that, the available quantity of degree-1 data is certainly insufficient.

from Briscoe's (2000), who observed constant oscillation in the population due to slight variation in the input to learners. The model here differs from his by using only unambiguous data to update the learner's hypothesis. I also allow the learner's final probability to be a value other than 0.0 or 1.0. I hypothesize that this is what yields the historically correct behavior. In addition, the model here has more realistic estimates for input quantity, population size, and learner lifespan.

4.6.3 The Feasibility of Filters

One might well be skeptical of the generality of the proposed filters. The unambiguous data filter in particular raises the question of how abundant such data points are for any given learning problem and the complexity of determining if a given data point is unambiguous. As a concrete example of both these issues for the word order case considered here, we can look to the "cartographic" approach to syntax (Rizzi, 2004; 1997; Cinque, 1999). This approach suggests that there are several positions in front of the VP that the Verb can move to if V2 movement is used. Languages are thought to differ on exactly which position it is. Given that, even knowing V2 movement has happened does not allow an unambiguous analysis of the sentence with respect to V2 movement; the learner still has more than one option for the Verb's exact position. If the initial intake is to contain any data points at all, it may be necessary to allow data points that are actually ambiguous to be *perceived* as unambiguous at the initial stages of learning.

If the learner is using cues to identify unambiguous data, then the level of specificity for a cue may be abstract enough to perceive ambiguous data as unambiguous. For instance, a cue may only specify one general position in front of the VP to identify V2 movement, rather than the multiple positions that the cartographic approach advocates. Only later would the learner then elaborate cues to include multiple positions in front of the VP. If the learner is using parsing to identify unambiguous data, then the learner could initially use a subset of the set of parameters an adult would use when parsing.[35] Later on, when more parameter values are known, the learner would expand the set of parameters used for parsing.

Another approach for both cues and parsing is that the learner has default values or assumptions (Fodor, 1998b) that are in place until the learner is forced to the marked values or assumptions. For example, in the word order case discussed here, the learner might assume as a default that there is no movement (thus perceiving simple SVO structures as unambiguous for VO word order). This assumption would then need to be revised at a later stage. The cost of reanalysis may not trivial, however, particularly when parameters and assumptions interact with each other.

Suppose, for instance, that default assumption A1 (e.g. no movement) allows the learner to perceive "unambiguous" data for a given value of P1 (e.g. OV/VO order), say, P1a (e.g. VO order). Later on, the learner is forced to remove default assumption A1. Suppose the lack of assumption A1 causes the learner to observe that (a) the "unambiguous" data for P1a are now ambiguous (e.g. SVO data) and (b) there now exist "unambiguous" data for P1b (e.g. more OV data). The learner must now

[35] A candidate set for the initial pool of parameters might be derived from a hierarchy of parameters, along the lines of the one based on cross-linguistic comparison that is described in Baker (2005, 2001).

re-evaluate the correct value for parameter P1 (OV/VO order), and so is delayed in attaining the adult target state. This same situation occurs when there are multiple parameters interacting (say, +/- V2 movement and OV/VO order). The issue of identifying unambiguous data in a system with multiple interacting parameters will be discussed in the next chapter.

The identification of unambiguous data is significantly aided by the assumption that parameters are independent structural pieces. Suppose we assume *n* parameters with 2 options each. If all parameters are independent, then every data point has at most 2*n* possible structural pieces that can be used to analyze it (Fodor, 1998a; 1998b; Sakas & Fodor, 1998). In contrast, if parameters are not independent, every data point can be analyzed with 2^n possible structures (since each "structure" is a combination of the smaller 2*n* structural pieces). It is thus enormously more efficient for ambiguity analysis to have independent parameters.

Moreover, if parameters are independent, data are unambiguous relative to a particular parameter. A given data point may be unambiguous for parameter P1 (e.g. OV ordering) while being ambiguous for many other parameters (e.g. wh-fronting). In contrast, if parameters are not independent, only data points that are unambiguous for *all* parameters are perceived as unambiguous – for otherwise, more than one structure of the available 2^n structural pieces leads to a successful analysis. Such data points are likely to be extremely sparse, if they exist at all.

4.6.4 Future Directions

Despite the ground covered in this chapter, there are of course a number of avenues that remain to be explored. The first concerns the relaxation of the unambiguous data filter, the second concerns the implementation of population models, and the third concerns experimental extensions.

In section 4.5.2.1, I explored one principled way a learner might use ambiguous data, which was to ignore possible movement rules in the system and assume that surface word order matched the underlying word order of the system. So, the hypothesis consistent with the surface order was fully credited for those data points, i.e. a data point with *Verb Object* anywhere in it would be credited to the Verb Object hypothesis. But there are other strategies that a learner might employ when encountering ambiguous data.

One method is to weight ambiguous data points such that they're not as influential as unambiguous data. In fact, I instantiated a method to do precisely this in the case study of anaphoric *one* in chapter 3, and the actual instantiation appears in the update procedure. The same concept of weighting could be applied to the syntax case examined in this chapter. If learners weight ambiguous data less than unambiguous data, it may be possible for them to achieve successful acquisition. If so, it behooves us to know what the successful weightings are for ambiguous and unambiguous data – and if we can find any experimental evidence to support such weightings.

Continuing the idea of weighting data, models of populations (such as the one examined here) can include additional sociolinguistic complexity in the relationships of the speakers that impact how learners view the data. Learners might, for instance, be more influenced by speakers who are in close spatial proximity, have a kinship relationship, or are from the same or higher social class. This weighting again would be instantiated in the update procedure. In addition, the frequency of various data types in the data intake distribution could depend on what speakers are nearby and/or are prominent in the learner's life. Family members will be a more frequent source of data than random, spatially distant population members.

Finally, the existence of data intake filtering for learning syntax – and specifically, using data perceived as unambiguous – can be explored in experimental regimes such as artificial language experiments for both adults (Thompson & Newport, 2007; Bonatti et al., 2005; Newport & Aslin, 2004) and children (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; among others). Specifically, learners could be exposed to data that would favor one word order if ambiguous data is used, but favor the other order if only unambiguous data is used. The generalization learners extract from such a dataset would implicate what data they use for learning.

4.6.5 Conclusion

In this chapter, I have investigated the effect of data intake filters in a system where the target adult state is a probability distribution between two opposing options for a single parameter. This was accomplished by employing language change modeling and using the assumption that a given case of language change was driven by language learning. Specifically, I adopted Lightfoot's (1991) assumption that Old English language change was driven by imperfect learning. The goal of the modeling was to see if I could replicate the precise amount of imperfect learning that causes the Old English population to change at a certain rate. As we have seen, this imperfect learning can result when the two data intake filters of unambiguous data and degree-0 data are used. Moreover, the historically correct population-level behavior does not result when either or both of the two filters is discarded, primarily because the data intake then does not have the proper bias in its distribution. Thus, through the language change model, I have provided empirical support for data intake filtering in language learning.

Now that we have seen evidence for the necessity of data intake filtering, we can now explore the feasibility of data intake filtering. This is particularly important for the unambiguous data filter, since identifying unambiguous data is a nontrivial task. In fact, it is quite reasonable to wonder how to identify unambiguous data in a system more complex than the one I have considered in this chapter (which considered 2 interacting parameters: OV/VO word order and +/- V2 movement). In the next chapter, I will examine the feasibility of the unambiguous data filter in a more complex system with 9 interacting parameters: English metrical phonology (Dresher, 1999).

Chapter 5: The Case of English Metrical Phonology

5.1 The Unambiguous Data Filter

We have just seen an argument for the necessity of an unambiguous data filter, using evidence from syntactic language change modeling. The motivation for the unambiguous data filter is that unambiguous data are the most informative data in a noisy data set; learning from informative data leads to convergence on the correct target state, given standard statistical learning techniques. So, unambiguous data are highly desirable since, once identified, they completely determine the choice in the hypothesis space for the learner.

Yet, I have also outlined why identifying unambiguous data is not a simple task. In particular, the identification of unambiguous data becomes considerably harder in a system with multiple interacting parameters. One might then wonder if it's even possible for an unambiguous data filter to be successful in a more complex domain, since this situation makes unambiguous data much sparser. In short, even if unambiguous data are desirable for learning, is it feasible to use an unambiguous data filter for a system with multiple interacting parameters?

For this reason, I turn now to the domain of metrical phonology, which has several interacting parameters, some of which have one or more sub-parameters, for a total of 9 interacting parameters.[36] Interacting parameters provide an additional challenge for language learners because the order in which these parameters are set by the learner (sometimes called the *learning path* (Dresher, 1999)) determines whether the learner will converge on the correct adult parameter values. If the choices are not made in the correct order, the child can misconverge. A reasonable question is whether we can discover *principled* metrics that allow the child to both find unambiguous data in the input and converge on an appropriate order of parameter-setting, given the noisy situation of multiple parameter interaction.

5.1.1 Two Methods For Identifying Unambiguous Data

First, we must ensure learner convergence on the adult system by uncovering the space of possible methods there are to discover sufficient unambiguous data in the correct distributions so that the learner converges on the adult system. Two methods have been proposed to implement an unambiguous data filter: a method that uses the domain-specific representation of cues (Dresher, 1999; Lightfoot, 1999 (see previous chapter)) and a method that uses the domain-specific learning procedure of parsing (Fodor, 1998b, 1998c; Sakas & Fodor, 2001).

A cue, according to Dresher (1999), is a "specific configuration in the input" that reflects a "fundamental property" of the particular parameter value it is a cue for. Moreover, a cue is local in the sense that a learner uses the cue "without regard to the

[36] This is much closer to the complexity that is purported to exist in the syntactic domain. A recent implementation by Sakas (2003) has 13 interacting parameters, though this is only a fraction of the parameters posited for the adult syntactic system.

final result", so that the learner is "not trying to match the input". The presence of a cue unambiguously favors one hypothesis (i.e., parameter value) in the hypothesis space over another.

The parsing method relies on the learner using the parsing strategies already available for language comprehension (Fodor, 1998b, 1998c; Sakas & Fodor, 2001). The learner tries to analyze a data point with "all possible parameter value combinations" in the hypothesis space given by universal grammar. The learner is, in effect, conducting an "exhaustive search of *all* parametric possibilities" (Fodor, 1998). If only a single parameter value for a given parameter is ever present in the successful parses for a particular data point, that data point is considered unambiguous for the parameter value. Data points that can be parsed with multiple parameter values are considered ambiguous. These ambiguous data points would then be filtered out of the learner's intake so that the update procedure is not activated when encountering them.

We will see that both these methods are successful at identifying sufficient unambiguous data to converge on the correct adult metrical phonology parameter values for English, a non-trivial task given the interactive nature of the parameters and the noise in the English data set. However, each method requires the learner to have different constraints on the order of parameter-setting.

5.1.2 Stipulation vs. Principled Derivation

Another relevant question is what pieces of each of these methods must be stipulated and what pieces can be derived in a principled fashion. As we will see, the cues method requires that we stipulate pre-specified knowledge in the form of the cues the learner uses to identify the unambiguous data. However, the constraints on the order of parameter-setting for the cues method result from properties of the learning system. In comparison, the parsing method does *not* need to stipulate extra information to identify unambiguous data – the process of assigning structure to input is already used for language comprehension. Yet, we will see that most of the constraints on the order of parameter-setting, in contrast to the cues method, must be stipulated.

5.1.3 Summary

In this chapter, I first explicitly compare the strengths and weaknesses of both the cues and parsing methods. Following that, I test both methods in the complex domain of metrical phonology, which has a set of 9 interacting parameters yielding 2^9 possible languages, and examine the potential of each method to identify sufficient unambiguous data to converge on the adult English parameter values. English is a particularly difficult case since there are actually unambiguous data in the input for the *incorrect* parameter values. Thus, the ability to converge on the correct parameter values in this exacerbated situation is a testament to the power of using either of these methods to identify unambiguous data. We will then see that both these methods can succeed and allow the learner to converge on the adult set of parameter values, providing support for the feasibility of an unambiguous data filter even in a more complex domain. I will then observe that the constraints on the order of parameter-setting that result from using each of these methods differ, and that the cues method allows us to derive the constraints in a principled manner while the parsing method requires that we stipulate the constraints. Finally, I will argue that both methods have strengths that can be combined and speculate that a more advantageous method of identifying unambiguous data comes from using a limited form of parsing to derive cues for unambiguous data.

5.2 The Cues Method and the Parsing Method

5.2.1 Cues: Strengths and Weaknesses

The cues method of finding unambiguous data has both strengths and weaknesses. Cues are attractive because they make identification of unambiguous data very simple: the data point either matches the cue, or it doesn't (though this, of course, assumes the learner can recognize the cue in the data point). In addition, a cue is designed to match a subpart of the data point, rather than the entire data point. This means the learner can glean information without understanding the structure for the entire data point. For instance, the learner can match a VO word order cue (example (1) taken from (8b) in the previous chapter) to a data point without understanding the structure for all the words in the sentence – the only words that are vital are the ones that correspond to the cue (Object, Verb, and some phrases that function as XP1 and XP2).

(1) VO word order cue: $[\ ]_{XP1}$ $[\ ]_{XP2}$ … Verb Object …

Because learners can extract information from only partial comprehension, cues offer a way to "get off the ground" when they don't know very much about the adult language.

Nonetheless, the cues method also has its weaknesses. Cues for each parameter are, by definition, a representation of domain-specific knowledge and must already be available for the learner or somehow derivable from previously available knowledge. In addition, the specificity of cues must be determined: are cues linear strings, underspecified structural pieces, or something else? Whatever the specificity of the cue, it must also be previously available for the learner or somehow derived. Thus, the cues method requires the learner to be equipped with additional knowledge (cues) to solve the language learning task.

Beyond this, some cues may require the learner to store data over time (perhaps in a summarized form) for comparison (Dresher, 1999). This is usually agreed to be an undesirable requirement in domains such as syntax because the potentially infinite number of sentences yield the possibility of unbounded storage requirements. However, the data storage requirement may fare better in domains such as metrical phonology where there are a finite number of morphemes and stress contours (even if in principle words can have infinitely many syllables).[37]

[37] Note, however, that the generative procedure for assigning stress can assign infinitely many stress contours in the same way that generative syntax can assign infinitely many structures. That is,

Another potential weakness is that cues are heuristic by nature, and so may lead to false positives or false negatives that could have a detrimental effect on learning over time.[38] For example, if we examine the cues for the OV/VO word order parameter ((1) and (2), taken from the previous chapter), we will notice that they only take V2 movement into account.

(2) OV word order cue: [$]_{XP}$ … Object Verb …

However, other grammatical rules in the adult language may also impact the observable data that these cues would match. Heavy Noun Phrase Shift is one such rule: it is a movement rule that shifts an Object that precedes a Verb to a position after the Verb, provided the Object is phonologically "heavy" enough.

As an abstract example, suppose the learner encounters a data point of the form "Adverb Subject Verb Object". This data point matches the VO cue "XP1 XP2 … Verb Object". Nonetheless, it could have been generated by starting with the order "Adverb Subject Object Verb" (which would have matched the Object Verb cue "XP …Object Verb") if the Object moved to a position after the Verb via Heavy Noun Phrase Shift. Since the observable data matches the learner's VO cue, the learner receives a false positive for VO order by using the heuristic cue.[39] If this kind of interference happens sufficiently, the learner may not converge on the correct adult value.

As a more concrete example of false positives from cues, consider the case of Kannada word order (data from Tirumalesh, 1996). The basic word order for Kannada is Object Verb (OV). A learner would encounter many examples of this order, as in (3a):

(3a) OV order in Kannada

raamu	dubai-ninda	**kumbaLakaayi**	**tand-id-d-aane**
Raam$_{\text{Subj}}$	Dubai-abl	**pumpkin$_{\text{Obj}}$**	**bring-be-npst-3sm$_{\text{Verb}}$**

'Raam has brought a pumpkin from Dubai.'

However, there is a rule in Kannada that will cause the observable order to have the Verb precede the Object (VO order), as in (3b). This rule applies when the meaning of the Object is surprising, and so the Object is moved after the Verb to put focus on its surprising nature.

generative procedures have no trouble coping with data of unbounded length (whether the data are words or sentences).

[38] Though this might help explain metrical stress change over time.

[39] This case also demonstrates how cues within the same language can conflict. If Heavy Noun Phrase Shift had not occurred because the Object was not "heavy" enough, the observed order would have been "Adverb Subject Object Verb." This order matches the OV cue. So, the very same language could have cues for both OV and VO word order. The learner presumably must then decide which is more likely to be the base order for the language, given the frequency of the different cues.

(3b) VO order in Kannada, due to surprise at ‘pumpkin’
raamu dubai-ninda **tandiddaane** **kumbaLakaayi**
Raam$_{\text{Subj}}$ Dubai-abl **bring-be-pst-3sm$_{\text{Verb}}$** **pumpkin$_{\text{Obj}}$**
"Raam has brought a pumpkin from Dubai.'

Since it is unusual to bring something as inexpensive as a pumpkin from Dubai, the Object ‘pumpkin’ would cause surprise in answer to a question like (3c).

(3c) Question in Kannada that would produce VO order
raamu dubai-ninda eenu tandiddaane
Raam Dubai-abl what bring-be-pst-3sm
'What has Raam brought from Dubai?'

Because the Object causes surprise, it is moved after the Verb to put focus on it. However, if a Kannada learner using cues is unaware of the rule that moves a surprising Object after the Verb, this learner might consider the data in (3b) as an example of VO order since it matches the cue “XP1 XP2 … Verb Object”. In this way, the learner receives a false positive for VO order in a language whose basic word order is OV. And again, if enough false positives (or false negatives) are encountered, the learner could fail to converge on the correct adult value of the given parameter.

Finally, some cues may require the learner to have default values among the options for a given parameter.[40] This means that the learner assumes that a given value holds unless there is evidence to the contrary. Note that the learner can still collect evidence to the contrary if data matches the cue for the non-default value, which is quite important if in fact the adult language uses the non-default value. Nonetheless, if default values are required for successful learning, these values are again an example of additional knowledge the learner requires specifically for solving the language learning task. After all, default values are representations of domain-specific knowledge that must be previously available or somehow derivable from previously available knowledge.

As an example of how the learner might derive a default value from previously available knowledge, suppose the candidate hypotheses are in a subset-superset relation, i.e. the set of data points that can be generated by one hypothesis are a subset of the set of data points that can be generated by the other hypothesis (as we saw in chapter 3 with anaphoric *one*). Under this viewpoint, the hypotheses are the opposing parameter values for the given parameter, which is knowledge the learner is assumed to already have available. The Subset Principle (Berwick,1985; Berwick & Weinberg, 1984) then provides the learner with a principled way to derive the default value: use the subset value.

[40] This could be instantiated as a hypothesis space with non-uniform prior probabilities. The initial probability distribution would be biased towards the default value.

5.2.2 Parsing: Strengths and Weaknesses

The parsing method for finding unambiguous data has its own strengths and weaknesses. One attractive feature is that parsing is a (domain-specific) procedure the learner already has available, assuming that the learner must come equipped with a procedure that tries to assign structure to the input given the available options (Fodor, 1998b, 1998c). In addition, the parsing method only requires one data point at a time, since it extracts as much information as possible and then proceeds to the next data point. No storage of data over time is required. Third, the parsing method, as discussed, is implemented as a find-all-parses analysis; it is therefore *not* heuristic. It will only find true unambiguous data, given the relevant parameter set. Finally, since all values are used during the find-all-parses analysis of the data point, no default values are required.

While this may seem like an impressive array of strengths, the parsing method also has its pitfalls. First, identification of unambiguous data is a non-trivial task requiring more resources from the learner, either in terms of multiple simultaneous parses stored in memory or in terms of using a sensible guessing strategy (Sakas & Fodor (2001) addresses the question of a sensible guessing strategy the learner might adopt for parsing). If the learner does a full find-all-parses analysis, we must explain how the learner can feasibly do this given finite resources; if a less resource-intensive guessing strategy is used, we must explain why the learner uses this strategy.

Second, if the entire data point cannot be parsed, no information can be extracted for *any* parameter. This makes “getting off the ground” during the initial stages of learning quite difficult, when the learner may not know enough to comprehend the entire data point (see Sakas & Fodor (2001), who acknowledge these problems and propose ways to solve them in scenarios where the adult language data does not contain numerous exceptions that lead to conflicting data points).

Beyond this, if exceptions exist in the input set that violate certain adult parameter values but obey others, those data points cannot be used by the learner since the learner cannot generate a successful parse of the data point.[41] In short, the parsing method does not allow information to be retrieved from subparts of a data point. One way to circumvent this problem would be for the learner to divide the data point into subparts using some sensible strategy (for instance, in syntax, the learner might divide a sentence into matrix and embedded clauses). Nonetheless, we would still need to provide a principled explanation for how the learner knows to divide up the data point in an appropriately helpful way.

Lastly, a learner using the parsing method may have difficulty finding unambiguous data if the relevant parameter set isn’t sufficiently restricted (too many possible parameters value sets could fit any given data point). This is perhaps best viewed as a problem of being too exacting about classifying data as unambiguous,

[41] If the learner can’t parse the data point, the learner presumably throws the entire data point out for the purposes of learning, classifying it as an exception that will have to be memorized in its entirety. Cues tolerate exceptions much more easily, allowing for anomalous sub-parts to be memorized instead of requiring the entire data point be memorized.

since the consideration of too many options would prohibit any data from being classified as unambiguous.

5.2.3 Summary: Cues vs. Parsing Overview

Both the cues and parsing methods have a large set of strengths and weaknesses, summarized in table 5.1 below. In this chapter, I explore an additional property for comparison: the effect each of these methods has on the learning path within the given domain of metrical phonology. Specifically, I examine the potential set of order constraints for parameter-setting that are generated by using each method to identify unambiguous data.

Property	**Cues**	**Parsing**
Easy identification of unambiguous data	True	False
Can get information from sub-part of data point	True	False
Can easily tolerate numerous exceptions in the data	True	False
Is heuristic	True	False
Requires additional prior knowledge for learner	True	False
Requires storage of data over time for comparison	True	False
Requires default values	True	False
Can work even in an unrestricted large set of initial parameters	True	False

Table 5.1. A comparison summary of the properties of the cues and parsing methods. Desirable properties are in **bold**, while potentially undesirable properties are in *italics*.

5.3 The Domain of Metrical Phonology

5.3.1 Why English Metrical Phonology?

The domain of metrical phonology has several merits for an investigation about the feasibility of unambiguous data (identified with either cues or parsing). First, although the parameter set consists of several parameters that interact in a complex fashion (Dresher, 1999), the set is small enough to make a find-all-parses approach more feasible and also provides a natural restriction on the relevant parameter set for the parsing method. In addition, though the parameter set is not as large as some implementations of syntax (Sakas (2003) implemented a version containing 13 interactive parameters[42]), it is still significantly more complex than the simplified case where the learner has only 1 or 2 interacting parameters to set.

Second, the cues method was originally proposed for this domain (Dresher, 1999), so there is some belief that it could be successful as an approach in general. It has also been used to study the acquisition of stress in English as a second language

[42] Note that 13 parameters is likely still a very small subset of the actual available syntactic parameters.

(Archibald, 1992). Third, the English system is not a toy example, and in fact is extremely messy. There are significant quantities of unambiguous data for *both values* of any given parameter. This makes the system non-trivial to learn because of the conflicting unambiguous data; the learner is required to extract systematicity from a very noisy environment.[43] The noisiness of the English data forces the learner to adopt order constraints on parameter-setting so that the *correct* systematicity is posited for the system. Because the parameters interact, it is easy for the learner to converge on the incorrect parameter value for a given parameter if the learner does not use order constraints.

The difficulty of this task makes the ability of either the cues method or the parsing method to learn the English system a major accomplishment already. The success of each of these methods lends support to the feasibility of an unambiguous data filter on the learner's intake, however such unambiguous data may be discovered by the learner.

5.3.2 Metrical Phonology Parameters

Metrical phonology is the system that determines which syllables in a word are stressed and how much stress each syllable receives compared to all the other syllables in the word. Here, I will be concerned with only the parameters that determine which syllables get stressed, and not with those which determine how much stress.

5.3.2.1 Parameters vs. Probabilistic Association

Given that there are a finite number of stress contours for words of *n* syllables, one might reasonably wonder why a parametric system is required instead of having the learner simply associate entire stress contours probabilistically with particular words (e.g. see Skinner (1957) for an associationist view of language learning, among many others). We can also translate this question to the realm of syntax: given that there aren't infinitely many parses for a given sentence, why don't we simply probabilistically associate structures with sentences, rather than having a procedure to generate these structures? Much ink has been spilled on this subject, with primary arguments coming from the finite syntactic variations across languages and the finite range of syntactic mistakes children make during learning.

One of the main arguments for a system of metrical phonology comes from stress change over time. Suppose learners simply associated stress contours probabilistically to individual words. Then, we would expect that stress change over time would proceed in a piece-meal fashion, with individual words changing at different times. Instead, we find cases where some historical linguists posit a swift change to an underlying *system* for analyzing stress contours that are assigned to words in order to best characterize the observed language change. This is because a number of words change at the same time, which would be quite coincidental if they were not somehow related. A very direct way to relate them is to say a common

[43] In fact, the English data are so messy that many linguists didn't believe English had any systematicity until Chomsky & Halle (1968).

system is used to generate their stress contours, and the change occurs to this system. Dresher & Lahiri (2003), for instance, note a particular shift in stress contours in Middle English between 1400 and 1530 (a relatively short time from a language change perspective) and posit a change to one parameter in the Middle English system in order to explain it.

Second, if there were no underlying system for generating the observed stress contours, we might also expect that when change occurs, the start and end states should be close to each other from a stress contour perspective. For instance, we might expect main stress to move from the final to the penultimate syllable. However, again we find examples where the start and end states do *not* seem closely linked with respect to the observable stress contours; instead, they are only close together when viewed in terms of a parametric system for generating the observed contours (again see Dresher & Lahiri (2003) for Middle English).

Because change can be sudden on a large scale and more easily explicable when viewed through the lens of a systematic representation for stress, it is believed that speakers represent stress contours in a systematic way that is richer than probabilistic association for individual lexical forms. Specifically for this chapter, I will assume speakers use the parameter system I outline below.

5.3.2.2 Parameters for Stress

I present a sketch of the metrical phonology parameters that are described more fully in Dresher (1999).[44] The parameter space is schematized in figure 31. Some parameters have only one level (e.g. Feet Headedness), while other parameters contain sub-parameters that become available if one option is chosen at the first level (e.g. Quantity Sensitivity).

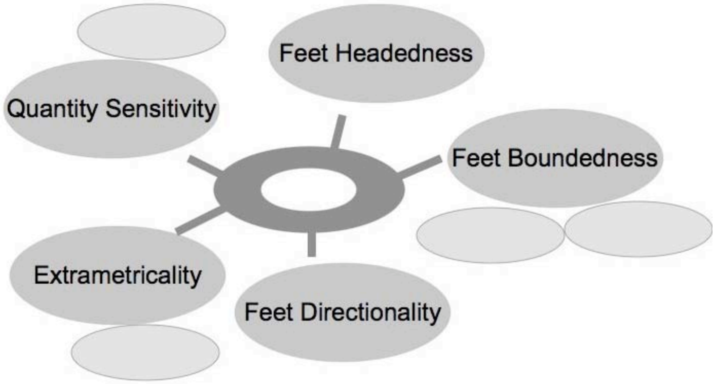


Figure 32. A schematic representation of the relevant parameters in metrical phonology, 5 main parameters and 4 subparameters for a total of 9 interacting parameters.

[44] Note that this parametric system differs from the instantiations in Halle & Idsardi (1995), Dresher (1994), and Idsardi (1992), though there are fairly straightforward mappings between the instantiation considered here and the ones considered in those studies.

A sample representation of metrical phonology structure is in figure 33.

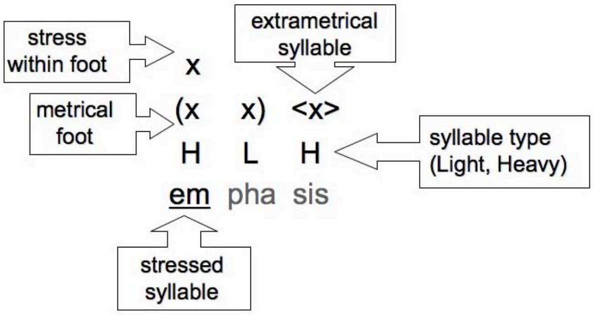


Figure 33. A sample representation of metrical phonology structure for ‘emphasis’, including terms to be described in more detail below: syllable type, metrical foot, extrametrical syllable, and stress within a metrical foot. In ‘emphasis’, the first and last syllables (‘em’, ‘sis’) are classified as Heavy, while the middle syllable (‘pha’) is classified as Light. The last syllable (‘sis’) is considered extrametrical, and not included in the metrical foot grouping. The first two syllables (‘em’, ‘pha’) are grouped into a single metrical foot, and the leftmost syllable in the foot (‘em’) is stressed.

5.3.2.2.1 Quantity Sensitivity

The first level of the quantity sensitivity parameter is whether the system is quantity-insensitive (QI) or quantity-sensitive (QS) (Halle & Idsardi, 1995; Hayes, 1980; among many others). An example language of this kind is Maranungku (Dresher, 1999). A quantity-insensitive system treats all syllables the same (represented as ‘S’ in (4)), whether they contain a long vowel as the nucleus (VV), a short vowel with a coda (VC), or a short vowel only (V). A long vowel syllable is “lu” in *ludicrous*, a short vowel with coda syllable is “crous” in *ludicrous* (the *s* is the consonant following the nucleus), and a short vowel only syllable is “di” in *ludicrous*. Note that the onset is irrelevant to syllable classification: VC, CVC, and CCVC are all classified as short vowels with codas and V, CV, and CCV are all classified as short vowels without codas. For syllables with a long vowel, the coda is also irrelevant – VV, VVC, and VVCC are all classified as long vowel syllables. In the examples below, all stressed syllables are underlined.

(4) ‘ludicrous’ analyzed in a QI system

syllable classification	**S**	**S**	**S**
nucleus & coda only	VV	V	VC
translation into V/C	CVV	CV	CCVC
syllables	<u>lu</u>	di	crous

A quantity-sensitive system divides syllables into (L)ight and (H)eavy[45]. Examples of this kind of language include Koya, Selkup, and Khalka Mongolian (Dresher, 1999; Halle & Idsardi, 1995; Hayes, 1980; among others). Long vowels (VV) are always Heavy while short vowels (V) are always Light. A subparameter then becomes available for how to classify short vowel syllables with codas (consonants following the vocalic nucleus (VC)), since some languages classify these as Light (VC-Light), e.g. Selkup (Dresher, 1999), while others classify them as Heavy (VC-Heavy), e.g. Koya (Dresher, 1999).

(5) 'ludicrous' analyzed in a QS system

(a) VC-Light

syllable classification	**H**	**L**	**L**
nucleus & coda only	VV	V	VC
translation into V/C	CVV	CV	CCVC
syllables	lu	di	crous

(b) VC-Heavy

syllable classification	**H**	**L**	**H**
nucleus & coda only	VV	V	VC
translation into V/C	CVV	CV	CCVC
syllables	lu	di	crous

Note that a syllable classified as H should have stress unless some other parameter interferes, such as extrametricality.

5.3.2.2.2 Extrametricality

Syllables in a word are grouped into larger units called metrical feet. Only syllables that are included in a metrical foot can be stressed. A syllable classified as extrametrical cannot be included in a metrical foot and therefore cannot receive stress. Only the syllable at the left or right edge of a word may be extrametrical, and only one syllable in the word may be extrametrical (both edge syllables cannot be extrametrical).[46]

A system can have no extrametricality (Em-None), so that all peripheral syllables are included in metrical feet. An example of this type of language is Maranungku (Dresher, 1999). Note that metrical feet are signified by parentheses (…) in the remaining examples.

[45] Though occasionally more complex weight systems have been proposed.

[46] There are additional proposed sub-classes of extrametricality that I will not consider here, such as (1) only Light edge syllables may be extrametrical (Hayes, 1980), (2) only the final syllable of nouns may be extrametrical (Hayes, 1980), (3) only the final consonant may be extrametrical (Archibald, 1998), and (4) only the final segment of the derivational stem (as indicated in the lexicon) can be extrametrical (Harris, 1983). Excluding these subparameters is an example of restricting the relevant parameter set for parsing.

(6) An Em-None analysis of ‘afternoon’, assuming QS-VC-Light; two metrical feet

syllable classification & metrical foot grouping	(L	L)	(H)
translation into V/C	VC	VC	VV
syllables	af	ter	noon

A system can also have extrametricality (Em-Some), e.g. English (Dresher, 1999), and then a subparameter becomes available to decide whether the leftmost syllable (Em-Left) or rightmost syllable (Em-Right) is the extrametrical one. Extrametrical syllables are signified by angle brackets <…> in the remaining examples.

(7) Em-Some analyses

(a) An Em-Left analysis of ‘agenda’, assuming QS-VC-Heavy; 1 metrical foot

syllable classification & metrical foot grouping	< L >	(H	L)
translation into V/C	V	VC	V
syllables	a	gen	da

(b) An Em-Right analysis of ‘ludicrous’, assuming QS-VC-Heavy; 1 metrical foot

syllable classification & metrical foot grouping	(H	L)	< H >
translation into V/C	VV	V	VC
syllables	lu	di	crous

As we can see in (7) above, the syllables that are classified as extrametrical do not receive stress. This is particularly striking in ‘ludicrous’, since the extrametrical syllable ‘crous’ is classified as Heavy under QS-VC-Heavy. Under normal circumstances, a Heavy syllable is usually stressed. Nonetheless, the extrametricality of the syllable interferes here, and allows the syllable to be without stress (and conform to the observed stress contour of ‘ludicrous’). This is one example of how different parameters can interact with each other.

5.3.2.2.3 Feet Directionality

Sequences of stressed syllables can be joined together as feet (Halle & Vergnaud, 1978; Hayes, 1995; Hayes, 1980; among many others). Metrical feet can be constructed beginning from the left side of the word (Ft Dir Left) or from the right side of the word (Ft Dir Right). An example of a language constructing feet from the left is Maranungku (Halle & Idsardi, 1995). Examples of languages constructing feet from the right are Warao and Weri (Halle & Idsardi, 1995).[47]

(8a) Metrical feet from the left (in a QS-VC-Light, Em-None system): L L H; 2 metrical feet

(i)	L	L	H
(ii)	(L	L	H
(iii)	(L	L)	H
(iv)	(L	L)	(H
(v)	(L	L)	(H)
Example stress contour:	L	L	H
Matching word:	pe	rox	ide

‘peroxide’

(8b) Metrical feet from the right (in a QS, Em-None system): L L H; 2 metrical feet

(i)	L	L	H
(ii)	L	L	H)
(iii)	L	(L	H)
(iv)	L)	(L	H)
(v)	(L)	(L	H)
Example stress contour:	L	L	H
Matching word:	ho	li	day

‘holiday’

As (8) shows, the syllables are divided differently into metrical feet, depending on the feet directionality. Since exactly one syllable in a metrical foot can receive stress, the differing metrical foot divisions can result in differing stress contours.

5.3.2.2.4 Boundedness

Boundedness refers to how large a metrical foot can be (Hayes, 1980; among many others). In an unbounded system (Unb), metrical feet can be arbitrarily large. The only reason a new metrical foot is started is if a Heavy syllable is encountered when grouping syllables into metrical feet. If, as in (9c) below, there are no Heavy syllables, then there will only be 1 metrical foot. Examples of this kind of language are Selkup and Koya (Dresher, 1999).

(9) Examples of unbounded analyses

(a) QS, Em-None, Ft Dir Left system: L L L H L; 2 metrical feet

(i)	L	L	L	H	L
(ii)	(L	L	L	H	L
(iii)	(L	L	L)	(H	L
(iv)	(L	L	L)	(H	L)

[47] Note that the examples below contain hypothetical analyses of the English words given as examples. In other words, those analyses are compatible with the stress contours observed.

(b) QS, Em-None, Ft Dir Right system: L L L H L; 2 metrical feet

(i)	L	L	L	H	L
(ii)	L	L	L	H	L)
(iii)	L	L	L	H)	(L)
(iv)	(L	L	L	H)	(L)

(c) QS, Em-None, Ft Dir Left system: L L L L L; 1 metrical foot

(i)	L	L	L	L	L
(ii)	(L	L	L	L	L
(iii)	(L	L	L	L	L)

In contrast, a bounded system places a limit on the size of the metrical foot, such that only a certain number of units are included. After that limit is reached, a new metrical foot is started. Examples of these kind of languages include Cayuvava, Warao, Weri, and Maranungku (Halle & Idsardi, 1995). Once the learner determines that the system is bounded, two subparameters become available: the size limit - 2 or 3 units (B-2 or B-3) - and what the counting units are - syllables or moras (B-Syl or B-Mor). Moras are units of syllable weight used in some languages (such as Japanese). If moras are the counting units, a Heavy syllable counts as two moras while a Light syllable counts as only one. Analyses using the various bounded options are in (10) and (11).

(10) Examples of bounded analyses: B-2 vs. B-3

(a) B-2, Em-None, Ft Dir Left: x x x x; 2 metrical feet

(i)	x	x	x	x
(ii)	(x	x	x	x
(iii)	(x	x)	x	x
(iv)	(x	x)	(x	x
(v)	(x	x)	(x	x)

(b) B-3, Em-None, Ft Dir Left: x x x x; 2 metrical feet

(i)	x	x	x	x
(ii)	(x	x	x	x
(iii)	(x	x	x)	x
(iv)	(x	x	x)	(x
(v)	(x	x	x)	(x)

(11) Examples of bounded analyses: B-Syl vs. B-Mor

(a1) QI, Em-None, Ft Dir Left, B-2, B-Syl: S S S S; 2 metrical feet

(i)	S	S	S	S
(ii)	(S	S	S	S
(iii)	(S	S)	S	S
(iv)	(S	S)	(S	S
(v)	(S	S)	(S	S)

(a2) QS, Em-None, Ft Dir Left, B-2, B-Syl: L H L L; 2 metrical feet

(i)	L	H	L	L
(ii)	(L	H	L	L
(iii)	(L	H)	L	L
(iv)	(L	H)	(L	L
(v)	(L	H)	(L	L)

(a3) QS, Em-None, Ft Dir Left, B-2, B-Syl: H H L L; 2 metrical feet

(i)	H	H	L	L
(ii)	(H	H	L	L
(iii)	(H	H)	L	L
(iv)	(H	H)	(L	L
(v)	(H	H)	(L	L)

(b) QS, Em-None, Ft Dir Left, B-2, B-Mor: H H L L; 3 metrical feet

(i)	H	H	L	L
	x x	x x	x	x
(ii)	H	H	L	L
	(x x	x x	x	x
(iii)	H	H	L	L
	(x x)	x x	x	x
(iv)	H	H	L	L
	(x x)	(x x	x	x
(v)	H	H	L	L
	(x x)	(x x)	x	x
(vi)	H	H	L	L
	(x x)	(x x)	(x	x
(vii)	H	H	L	L
	(x x)	(x x)	(x	x)
(viii)	H	H	L	L
	(x x)	(x x)	(x	x)
(ix)	(H)	(H)	(L	L)

As (11a3) and (11b) demonstrate, using syllables instead of moras as the counting units can create a markedly different metrical foot structure, which then affects the observed stress contour.

5.3.2.2.5 Feet Headedness

Feet headedness refers to which syllable in a metrical foot receives stress – the leftmost (Ft Hd Left) or the rightmost (Ft Hd Right) (Hayes, 1980; among many others).

(12) Examples of analyses with Ft Hd Left and Ft Hd Right – stressed syllables

underlined
(a) QI, Em-None, Ft Dir Left, B-2, B-Syl, **Ft Hd Left**: S S S
($\underline{S}$ S) ($\underline{S}$) → $\underline{S}$ S $\underline{S}$
(b) QI, Em-None, Ft Dir Left, B-2, B-Syl, **Ft Hd Right**: S S S
(S $\underline{S}$) ($\underline{S}$) → S $\underline{S}$ $\underline{S}$

5.3.2.3 Interacting Parameters

As we can see, all the metrical phonology parameters interact in their effect on the final stress contour assigned to a given word; a change to any one of them could change the stress contour in a non-trivial fashion. An example is illustrated in (13): the change of one parameter value (Em-None to Em-Left) causes the entire stress contour to become its inverse.

(13) A change to one parameter can drastically affect the stress contour assigned
(a) QI, **Em-None**, Ft Dir Left, B-2, B-Syl, Ft Hd Left: S S S S S
3 metrical feet
($\underline{S}$ S) ($\underline{S}$ S) ($\underline{S}$) → $\underline{S}$ S $\underline{S}$ S $\underline{S}$
Example: Maranungku 'langkaratati' → $\underline{\text{lang}}$ ka $\underline{\text{ra}}$ ta $\underline{\text{ti}}$

(b) QI, **Em-Left**, Ft Dir Left, B-2, B-Syl, Ft Hd Left: S S S S S
2 metrical feet
< S > ($\underline{S}$ S) ($\underline{S}$ S) → S $\underline{S}$ S $\underline{S}$ S
Example: Maranungku 'langkaratati' – incorrect stress pattern
→ lang $\underline{\text{ka}}$ ra $\underline{\text{ta}}$ ti
Example: English 'communication' – correct stress pattern
→ co $\underline{\text{mmu}}$ ni $\underline{\text{ca}}$ tion

Moreover, ambiguity can also easily arise – a single stress contour could be covered by multiple combinations of different parameter values, as shown in (14). Note that these combinations yield identical stress contours for 'communication', but these combinations may well yield differing stress contours for other words. Thus, the collection of combinations that produce the observable stress contour for any given word will vary from word to word.

(14) Multiple analyses of a single stress contour:
some analyses of 'communication' = co $\underline{\text{mmu}}$ ni $\underline{\text{ca}}$ tion
(a) QI, Em-Left, Ft Dir Left, B-2, B-Syl, Ft Hd Left
2 metrical feet
< S > ($\underline{S}$ S) ($\underline{S}$ S) → S $\underline{S}$ S $\underline{S}$ S
(b) QI, Em-Right, Ft Dir Left, B-2, B-Syl, Ft Hd Right
2 metrical feet
(S $\underline{S}$) (S $\underline{S}$) < S > → S $\underline{S}$ S $\underline{S}$ S

(c) QS, QSVCH, Em-Right, Ft Dir Right, B-2, B-Syl, Ft Hd Right

2 metrical feet
(L $\underline{H}$) (L $\underline{H}$) < H > → L $\underline{H}$ L $\underline{H}$ L
(d) QS, QSVCH, Em-Right, Ft Dir Left, B-3, B-Mor, Ft Hd Right
2 metrical feet
(x x x) (x x x) x x
(L $\underline{H}$) (L $\underline{H}$) < H > → L $\underline{H}$ L $\underline{H}$ L

Converging on the correct values for the adult system with its interacting parameters is thus not a simple task. Because the parameters all combine to produce the observable stress contour, identifying unambiguous data for a *single* parameter value is not easy. Nonetheless, this is precisely what the cues and parsing methods are proposed to do. I will now describe how both methods would identify unambiguous data for each of the values of each of these parameters, thereby instantiating the unambiguous data filter on the learner's intake.

5.4 The Cues Method for Finding Unambiguous Data

The cues method makes identification of unambiguous data simple, provided the learner knows the relevant cues and can match them to the data encountered. Recall that the cues method was originally proposed by Dresher (1999) for the metrical phonology domain, and he described a set of potential cues for each of the parameters. One property of his cue set is that it assumes some parameters values are the default, and cues are only for the marked values. As I noted previously, this could be perceived as a pitfall since it requires the learner to have pre-specified domain-specific knowledge (perhaps as a non-uniform prior probability distribution biased towards the default value). Dresher (1999) suggests a way to derive this knowledge: learners begin with simple representations and must be driven to more complex representations (in the spirit of Chomsky & Halle (1968)). He proposes that his default values are simpler representations than their marked counterparts.

Nonetheless, since default values are an additional stipulation about the learner's knowledge, I provide an alternate set of cues that does not require defaults; each opposing parameter value has its own cue. I will compare the performance of these two cue implementations on the metrical phonology data.

5.4.1 Quantity Sensitivity

In the cue set proposed by Dresher (1999), the value where the syllables are undifferentiated (QI) is the default value. The cue for QS (where the syllables are classified as either Light and Heavy) is to compare words with the same number of syllables. If they have different stress contours, then the system is QS.

(15) Dresher cues for quantity sensitivity
(a) QI: default value (no cue required)
(b) QS: 2 words with *n* syllables that have different stress contours
Ex: $n = 2$, word 1: $\underline{VV}$ V, word 2: $\underline{VV}$ $\underline{VV}$

An alternate cue set has cues for both QI and QS, as well as for the subparameters of QS (QS-VC-Light, where a VC syllable is treated as Light, and QS-VC-Heavy, where a VC syllable is treated as Heavy). The cue for QI is to find an unstressed internal VV syllable (which would be Heavy in a QS system, and therefore likely to be stressed) (16a). The cue for QS is to find a 2 syllable word with 2 stresses (or a 3 syllable word with 2 adjacent stresses if the system is known to be extrametrical already) (16b). Once the system is known to be QS, the cue for QS-VC-Light is an unstressed internal VC syllable (16c) while the cue for QS-VC-Heavy is a 2 syllable word with 2 stresses, where at least one syllable is VC (or the 3 syllable variant if extrametricality is known to apply) (16d).

(16) Alternate cues for quantity sensitivity

(a) QI: unstressed internal VV syllable

Ex: VV VV VV

(b) QS: 2 syllable word with 2 stresses, or 3 syllable word with 2 adjacent stresses if extrametricality is known

Ex: (1) VV VV (2) Em-Right: VV VV VV

(c) QS-VC-Light: unstressed internal VC syllable

Ex: VV VC VV

(d) QS-VC-Heavy: 2 syllable word with 2 stresses, with at least one syllable VC (or 3 syllable word with 2 adjacent stresses and at least one syllable VC if extrametricality is known)

Ex: (1) VV VC (2) Em-Right: VV VC VV

Note that if a default-marked system was preferred, the QI and QS-VC-Light values would function as the default values, with cues existing for QS and QS-VC-Heavy. I offer some speculation as to why the QI and QS-VC-Light values might be the default. One could argue that a QI system, because it treats all the syllables as the same, is a simpler method than dividing syllables into Light and Heavy. One could also argue that a QS-VC-Light system is simpler than a QS-VC-Heavy system. In particular, if a division between Light and Heavy syllables must be made, and Heavy syllables are marked in some way, having only VV syllables be Heavy is simpler than having other syllables such as VC also be Heavy.

5.4.2 Extrametricality

In the cue set proposed by Dresher, having an extrametrical syllable is the default state. This may be a difficult default to defend, however, since one might view extrametricality (i.e. ignoring certain edge syllables) as a marked feature of the metrical structure that the learner would need evidence for. Nonetheless, in the Dresher (1999) system, cues rule out extrametricality for each side (Em-Left and Em-Right). To rule out extrametricality for a given side, the edge syllable (leftmost for Em-Left and rightmost for Em-Right) must have stress.

(17) Dresher cues for extrametricality

(a) Em-None: Both leftmost and rightmost syllables have stress
Ex: $\underline{\text{VV}}$ $\underline{\text{VC}}$
(b) Em-Some (Left or Right): default

An alternate cue set has cues for Em-Some (both Em-Left and Em-Right) as well as for Em-None. The cue for no extrametricality (Em-None) is similar to the Dresher-style cue: both edge syllables are stressed (18a). The cue for Em-Some is that a Heavy syllable at either edge of the word is unstressed (18b); the cue for Em-Left is that the leftmost syllable is Heavy and unstressed (18c) while the cue for Em-Right is the rightmost syllable is Heavy and unstressed (18d).

(18) Alternate cue set for extrametricality
(a) Em-None: Both leftmost and rightmost syllables have stress
Ex: $\underline{\text{VV}}$ $\underline{\text{VC}}$
(b) Em-Some: Either edge syllable is Heavy and unstressed
Ex: (1) H L $\underline{\text{H}}$ (2) $\underline{\text{H}}$ L H
(c) Em-Left: Leftmost syllable is Heavy and unstressed
Ex: H L $\underline{\text{H}}$
(d) Em-Right: Rightmost syllable is Heavy and unstressed
Ex: $\underline{\text{H}}$ L H

Note again that the alternate cue set could also be set up as a default-marked system. In the alternate cue set, having no extrametricality (Em-None) could be argued as the default under the assumption that all syllables should be included for metrical feet groupings until the learner is forced by evidence to do otherwise.

5.4.3 Feet Directionality

The cue set proposed by Dresher requires the feet directionality cues to be combined with the feet headedness cues, and so I will examine these cues together in section 5.4.5. An alternate cue set has cues for feet directionality separate from cues for feet headedness.

In the alternate set, the cue for Feet Directionality Left is dependent on the quantity sensitivity value. If the system is quantity insensitive (QI), the cue is 2 stressed adjacent syllables at the right edge of the word (19a1); if the system is quantity sensitive (QS), the cue is 2 stressed adjacent syllables with the first syllable Heavy and the second Light at the right edge of the word (19a2). In addition, if the system is known to have extrametricality on the rightmost syllable, then the cue is shifted to the previous two syllables. The cue for Feet Directionality Right is exactly the same, except that the 2 stressed adjacent syllables are at the left edge of the word (19b).

(19) Alternate cue set for feet directionality

(a) Feet Directionality Left

(1) If QI: 2 stressed adjacent syllables at the right edge of the word (if extrametricality exists for the rightmost syllable, the 2 stressed adjacent syllables are shifted over one position)
Ex: (1) S S $\underline{S}$ $\underline{S}$ (2) S S $\underline{S}$ $\underline{S}$ <S>

(2) If QS: 2 stressed adjacent syllables at the right edge of the word, with the first as H and the second as L (if extrametricality exists for the rightmost syllable, the 2 stressed adjacent syllables are shifted over one position)
Ex: (1) L L $\underline{H}$ $\underline{L}$ (2) L L $\underline{H}$ $\underline{L}$ <L>

(b) Feet Directionality Right

(1) If QI: 2 stressed adjacent syllables at the left edge of the word (if extrametricality exists for the leftmost syllable, the 2 stressed adjacent syllables are shifted over one position)
Ex: (1) $\underline{S}$ $\underline{S}$ S S (2) <S> $\underline{S}$ $\underline{S}$ S S

(2) If QS: 2 stressed adjacent syllables at the left edge of the word, with the first as L and the second as H (if extrametricality exists for the leftmost syllable, the 2 stressed adjacent syllables are shifted over one position)
Ex: (1) $\underline{L}$ $\underline{H}$ L L (2) <L> $\underline{L}$ $\underline{H}$ L L

5.4.4 Boundedness

In the cue set proposed by Dresher, the Unbounded value is the default and cues signal that the system is bounded. The cue for boundedness is the presence of an internal stressed Light syllable.

(20) Dresher cues for boundedness
(a) Unbounded: default
(b) Bounded: an internal stressed Light syllable
Ex: L L $\underline{L}$ L

An alternate cue set has cues for both Unbounded and Bounded, as well as for the subparameters of Bounded (B-2 vs. B-3, B-Syl vs. B-Mor). The cue for an unbounded system depends on the system's quantity sensitivity. If the system is QI, the cue is three or more unstressed syllables in a row (21a1); if the system is QS, the cue is three or more unstressed Light syllables in a row (21a2).[48]

The cue for a bounded system is really the union of the cues for B-2 and B-3, which are again dependent on the quantity sensitivity of the system. If the system is QI, the B-2 cue is three or more syllables in a row with every other syllable stressed

[48] Note that this cue can interact with extrametricality. If the learner knows the system is extrametrical (either left or right), that syllable would be excluded from the three (or more) unstressed syllables necessary to be an Unbounded cue.

(21c1); if the system is QS, the cue is three or more Light syllables in a row with every other syllable stressed (21c2). The B-3 cue is nearly identical, except that there must be four or more (Light) syllables in a row with every third one stressed (21d).

This leaves the cues for a system that counts syllables (B-Syl) vs. a system that counts moras (B-Mor). The B-Syl cue also depends on the quantity sensitivity of the system. If the system is QI, then the cue is identical to the B-2 and B-3 cues (3+ syllables in a row with every other one stressed or 4+ syllables in a row with every third one stressed) (21e1). If the system is QS and B-2, the cue is 2 adjacent syllables with the pattern ‘$\underline{H}$ L’ or ‘L $\underline{H}$’ (21e2); if the system is QS and Bounded-3, the cue is 3 adjacent syllables with the pattern ‘$\underline{H}$ L L’ or ‘L L $\underline{H}$’ (21e3).

The B-Mor cue is far simpler: a 2 syllable word with both syllables stressed, and both syllables are Heavy (21f). If extrametricality is known to apply, then the cue is the same except that it applies to the 2 adjacent syllables that aren’t extrametrical.

(21) Alternate cue set for boundedness

(a) Unbounded: 3+ unstressed (Light) syllables in a row

Ex: (1) QI: S S S $\underline{S}$ (2) QS: L L L $\underline{L}$

(b) Bounded: union of B-2 and B-3 cues

(c) B-2: 3+ (Light) syllables in a row, every other one stressed

Ex: (1) QI: $\underline{S}$ S $\underline{S}$ S (2) QS: $\underline{L}$ L $\underline{L}$ L

(d) B-3: 4+ (Light) syllables in a row, every third one stressed

Ex: (1) QI: $\underline{S}$ S S $\underline{S}$ (2) QS: $\underline{L}$ L L $\underline{L}$

(e) B-Syl:

(1) QI: is union of B-2 and B-3 cues for QI

(2) QS, B-2: 2 adjacent syllables with pattern ‘$\underline{H}$ L’ or ‘L $\underline{H}$’

Ex: (1) $\underline{L}$ L $\underline{\mathbf{H}}$ **L** (2) **L** $\underline{\mathbf{H}}$ L $\underline{L}$

(3) QS, B-3: 3 adjacent syllables with pattern ‘$\underline{H}$ L L’ or ‘L L $\underline{H}$’

Ex: (1) $\underline{\mathbf{H}}$ **L L** $\underline{L}$ L (2) L $\underline{L}$ **L L** $\underline{\mathbf{H}}$

(f) B-Mor: 2 syllable word with both syllables stressed and Heavy

Ex: (1) $\underline{H}$ $\underline{H}$ (2) (Em-Left) < L > $\underline{H}$ $\underline{H}$

The complexity of some of the cues for boundedness suggests that a default-marked system might be quite attractive here. In this case, I speculate that Unbounded would be the default, since it is an assumption that there is no arbitrary metrical foot size.[49] Also, counting by syllables (B-Syl) as opposed to moras (B-Mor) could be argued as the default, since words are already divided into syllables for many of the other parameters.

5.4.5 Feet Headedness

The set of cues proposed by Dresher has a single “cue” for feet directionality and feet headedness. In fact, this cue is really very much like a find-all-parses analysis using the restricted parameter set F = {Feet Directionality, Feet

[49] Also, the Unbounded value as default falls out from the metrical phonology system implemented by Idsardi (1992).

Headedness}. The learner parses the known set of words with all combinations of feet headedness and feet directionality ((1) Ft Dir Left/Ft Hd Left, (2) Ft Dir Left/Ft Hd Right, (3) Ft Dir Right/Ft Hd Left, (4) Ft Dir Right/Ft Hd Right). For a given combination, if all the known words can be parsed such that all Light syllables that aren't the head of a metrical foot are unstressed, then this situation is the "cue" for this combination of feet directionality and feet headedness.

However, as I proposed an alternate set of cues for feet directionality by itself, I propose an alternate set for feet headedness by itself. The cue for Feet Headedness Left is that the leftmost syllable of the leftmost foot is stressed (22a); the cue for Feet Headedness Right is that the rightmost syllable of the rightmost foot is stressed (22b).

(22) Alternate cue set for feet headedness[50]
(a) Feet Hd Left: the leftmost syllable of the leftmost foot is stressed
(1) VV VC V (2) (Em-Left) < VV > VC V V
(b) Feet Hd Right: the rightmost syllable of the rightmost foot is stressed
(1) VV VC V (2) (Em-Right) VC V VC < VV >

5.4.6 Summary: Cues

I have now stepped through cues for each of the relevant parameters in the metrical phonology domain. Recall that one of the strengths of cues is that the learner can easily identify unambiguous data, since it will match the cue the learner knows. However, the cues proposed here are heuristic in nature and may cause the learner to perceive false positives and false negatives, which could in turn lead the learner astray.

5.5 The Parsing Method for Finding Unambiguous Data

5.5.1 The find-all-parses analysis

The parsing method differs from the cues method in that it is more resource-intensive to identify unambiguous data, but far less likely to identify false positives and false negatives. A learner using the parsing method will parse the given data point with all *available* values of all the parameters in the relevant parameter set. Note that a parameter value ceases to be available when the learner decides the other value is correct for the language. For instance, if the learner has decided that the system is QS, no parses will be generated that use the value QI.

I have termed this procedure the find-all-parses analysis. While there are other implementations of the parsing method that are not as resource-taxing as the find-all-parses analysis, the find-all-parses analysis is the most inclusive version. I want to give the parsing method the best chance for successful identification of

[50] The feet headedness cues can actually apply to the same word if it has stress on both the initial and final syllable. However, the learner effectively learns nothing from such a word since neither parameter value gets the advantage over the other from this word. Alternatively, the learner might choose to explicitly ignore such a word as inconsistent, since it displays cues for mutually exclusive parameter values.

unambiguous data to see how it compares to the cues method. If this version of the parsing method is superior to the cues method, then we can see if weaker versions using less resources are also superior. First, however, I investigate whether the find-all-parses implementation can get the job done.

After the learner has conducted a find-all-parses analysis on the data point, the learner then sees if only one parameter value of a parameter leads to a successful parse of the data point. If so, the data point is considered unambiguous for that value. The results of an example find-all-parses analysis for a data point are shown in (23).

(23) The results of a find-all-parses analysis of ‘<u>af</u> ter <u>noon</u>’: sets of parameter values that yield a matching stress contour

(a) (QI, **Em-None**, Ft Dir Left, B, B-2, B-Syl, Ft Hd Left)
(b) (QI, **Em-None**, Ft Dir Rt, B, B-2, B-Syl, Ft Hd Rt)
(c) (QS, QS-VCL, **Em-None**, Ft Dir Left, Unb, Ft Hd Left)
(d) (QS, QS-VCL, **Em-None**, Ft Dir Left, B, B-2, B-Syl, Ft Hd Left)
(e) (QS, QS-VCL, **Em-None**, Ft Dir Rt, B, B-2, B-Syl, Ft Hd Rt)

Since all successful parses share the parameter value Em-None, this data point would be considered unambiguous for Em-None.

Note that if the relevant parameter set is restricted, the find-all-parses analysis returns fewer parameter value sets and the same data point may then be considered unambiguous for other parameter values. Thus, a data point may be viewed as unambiguous for *different* parameter values at different points in time. This emphasizes how the definition of “unambiguous” is relative to the learner’s current knowledge state. This shift in perceived “unambiguity” is demonstrated in (24). In this example, suppose the learner has determined that the system is QI. The find-all-parses analysis will disregard any parses that include QS values.

(24) The results of a find-all-parses analysis of ‘<u>af</u> ter <u>noon</u>’, with the restriction that the system is QI: sets of parameter values that yield a matching stress contour

(a) (QI, **Em-None**, Ft Dir Left, **B, B-2, B-Syl**, Ft Hd Left)
(b) (QI, **Em-None**, Ft Dir Rt, **B, B-2, B-Syl**, Ft Hd Rt)

Given these results from the find-all-parses analysis, the learner using the parsing method would view this data point as unambiguous for Em-None, Bounded, Bounded-2, and Bounded-Syl.

5.5.2 Summary: Parsing

I have now described how the parsing method identifies unambiguous data in the input. Recall that one of the strengths of parsing is that it is not heuristic in nature and therefore will not perceive false positives or false negatives. The simplicity of the method is also appealing, since the learner needs only to use a procedure already available for language comprehension.

However, the find-all-parses method proposed here has its own pitfalls. I reiterate that it may be quite resource-intensive for a learner to implement. Moreover, a fundamental problem with the parsing method is that it cannot use data it cannot parse. In a language learning situation in which there aren't many exceptions to the adult parameter values, this is not too damaging. But the English data set, as I have mentioned before, is fraught with such exceptions. Once the learner has set some of the parameter values correctly, it will be unable to parse the non-trivial portion of the data that are exceptions. In this sense, the parsing method may not be "flexible" enough to cope with noisy data. We will see how this inflexibility impacts the learning path the learner must take to converge on the correct set of parameter values for English. First, however, we will verify the performance of these two methods on an easier language learning case where the available data set is not exception-filled.

5.6 Cues and Parsing in a Clean Language Environment: An Easy Case

A "clean" language environment is one in which there are *no* conflicting unambiguous data in the learner's input. A clean language environment makes convergence on the adult parameter values very straightforward. Given sufficient time, the learner will be exposed to enough unambiguous data points to converge on the adult parameter values. There are no "garden paths" provided by unambiguous data for the incorrect parameter values (in contrast to noisy data sets such as English). As long as these methods allow the learner to perceive *some* data as unambiguous, the learner will eventually converge on the correct parameter values. I sketch how these methods would work for a language like Maranungku, which has stress on every odd syllable counting from the left (examples in (25) from Dresher (1999) and Kager (1995)).

(25) Maranungku Stress Contour Examples (stressed syllables underlined)

(a)	lang	ka	ra	ta	ti		- 'prawn'
(b)	we	le	pe	le	man	ta	- 'kind of duck'
(c)	ya	ngar	ma	ta			- 'the Pleiades'
(d)	me	re	pet				- 'beard'
(e)	ti	ralk					- 'saliva'

The parameter values for Maranungku are in (26).

(26) Maranungku metrical phonology parameter values
(QI, Em-None, Ft Dir Left, B, B-2, B-Syl, Feet Hd Left)

Each of the words in (25) can be analyzed with either the cues or parsing method to identify if any are unambiguous for any of the parameter values (see table 5.2). As we can see, the two methods do not always agree on how many values a given data point is unambiguous for. Nonetheless, all the data are always unambiguous for the correct adult system parameter values, due to the clean language environment.

	Unambiguous for Cues	Unambiguous for Parsing

lang ka ra te ti	Bounded-2, Em-None	Bounded-2, Em-None
we le pe le man ta	Bounded-2, Ft Hd Left	Bounded-2
ya ngar ma ta	Ft Hd Left	Bounded-2
me re pet	Em-None	Em-None
ti ralk	Ft Hd Left	*Nothing*

Table 5.2. The results of using the cues and parsing methods to classify five Maranungku words as unambiguous for available parameter values.

Because there are no conflicting unambiguous data, neither the cues nor parsing method would classify the data as unambiguous for the incorrect parameter value. The learner should thus converge on the adult Maranungku system no matter which method is used, given exposure to sufficient data. Moreover, there are no order constraints on the learning path: the learner should converge on the correct adult values, no matter what order the learner sets the parameters in.

5.7 Learning English: A Harder Case

English, however, poses a more difficult challenge since it *does* have conflicting unambiguous data points, as perceived by the learner. I turn now to how I tested each of these methods for learning the English metrical phonology system.

5.7.1 Estimating the Composition of the Input to the Learner

I compiled caretaker speech to children between the ages of 6 months and 2 years from the CHILDES database (MacWhinney, 2000), for a total of 540505 words. Each of these words were then divided into syllables and marked with stress, using the CALLHOME database of telephone conversation (Canavan et al., 1997) and the MRC psycholinguistics database (Wilson, 1988) as references for likely syllabic divisions and stress contours. I assumed that this was a reasonable estimation of the composition of the data English learners would be exposed to.

5.7.2 The Logical Problem of Learning English Metrical Phonology

The correct parameter values for English are listed in (27).

(27) English metrical phonology parameter values
(QS, QS-VC-Heavy, Em-Right, Ft Dir Right, Ft Hd Left, B, B-2, B-Syl)

Converging on the correct parameter values for English adults is non-trivial, given realistic distributions of input to English children. We must ask what parameter-setting orders (if any) will lead the learner to converge on the adult parameter values. Importantly, every time learners set one parameter, they may then view all subsequent data differently. So, the setting of one parameter in one way could bias the learner to set another parameter in another way later on. Thus, the order of parameter-setting can have a significant effect on the final set of parameter values the learner converges on. A viable parameter-setting order will lead the learner to converge on the correct set of parameter values for the language.

The viable orders are derived via an exhaustive walk through all possible parameter-setting orders; hence, this is exploring the logical problem of learning, in that we are interested in whether the target state is achievable at all using these learning methods. To conduct the exhaustive walk for a given learning method (cues or parsing), we must try out every single parameter-setting order with the input.

In the worst case, no order will suffice – the target set of parameter values is unreachable, given the input and this learning method. Learning with an unambiguous data filter produces insufficient behavior.

A better scenario is that learning with an unambiguous data filter *does* produce sufficient behavior. A slightly better case is that there *is* a set of orders that will allow the learner to reach the target set, but these orders are completely unrelated to each other. There is no way to make the knowledge necessary for acquisition success concise; the learner must somehow be aware of the viable orders explicitly. In an even better case, there is a set of orders that will work, and they can be captured by a small number of *order constraints*, though these order constraints may need to be stipulated. A still better case is that a set of viable orders exists that can be captured by *principled* order constraints that are independently derivable. In the best case, all parameter-setting orders will be viable so there is no need to worry about the order of parameter-setting at all.[51] In this last case, since there are no constraints on the order of parameter-setting, there is no need to explain how the learner knows them or why the learner follows them.

5.3 Conducting an Exhaustive Walk Through All Possible Orders

5.3.1 The Algorithm for Identifying All Viable Parameter-Setting Orders

Here, I describe the method for conducting an exhaustive walk through all possible parameter-setting orders to determine which, if any, will lead the learner to converge on the adult set of parameter values.

(28) Algorithm for identifying all viable orders of parameter-setting for a given learning method

(a) For all currently unset parameters, determine the unambiguous data distribution in the corpus (i.e. how much unambiguous data there is for each value of each unset parameter).

(b) Choose a currently unset parameter to set. The value chosen for this parameter is the value that has a higher probability in the data the learner perceives as unambiguous. This logic behind this is that, given enough data points (i.e. a sufficiently long learning period), this parameter value will eventually accrue enough probability to become the winning parameter value.

(c) Repeat steps (a-b) until all parameters are set.

(d) Compare final set of values to target set of values. If they match, this is a viable parameter-setting order.

[51] This is the case of the clean language environment described in the previous section.

(e) Repeat (a-d) for all parameter-setting orders.

The process of determining the distribution of unambiguous data in this corpus after each parameter is set (28a) is meant to reflect how the learner perceives the incoming data at different points in the parameter-setting process. I want to use all the data available in the sample corpus to estimate what the input distributions are for the learner at any given point in time. Thus, after each parameter is set, this algorithm gauges how the learner would then view the available input in the linguistic environment by recalculating the unambiguous data distributions in the corpus.

In (28b), the learner chooses the parameter value with a higher probability in the unambiguous data. There are two ways to measure unambiguous data probability, depending on what the learner is relativizing the probability against. One way, which I will refer to as the *relativize-against-all* approach, relativizes the unambiguous data for that parameter value against the entire input set. The second way, which I will refer to as the *relativize-against-potential* approach, is for the learner to relativize the unambiguous data for that parameter value against the set of *potential* unambiguous data points. The set of potential unambiguous data points is smaller than the entire input set and may vary across parameters, since not every data point satisfies the preconditions necessary to be an unambiguous data point. Moreover, the preconditions will vary depending on whether the learner uses cues or parsing to identify unambiguous data. I will describe in detail below why this occurs. Meanwhile, it is unclear a priori which relativization approach should be preferred by the learner, so I will examine the effects of both separately.

5.7.3.2 Relativization of Unambiguous Data Probability

The relativize-against-all approach can intuitively be characterized by the question, “How likely is it that a random data point chosen from the entire input set will be an unambiguous data point for the parameter value of interest?” It does not matter for this approach whether unambiguous data are identified via cues or via parsing because the relativizing set (the input set size) is constant across both cues and parsing.

As a concrete example, suppose the data set provides 11213 data points perceived by the learner as unambiguous for Quantity Sensitive (QS) and 2140 perceived as unambiguous for Quantity Insensitive (QI) . The total data set size is 540505 words, so the relativized probability for an unambiguous QS data point is 11213/540505 = 0.0207 and the relativized probability for an unambiguous QI data point is 2140/540505 = .00396. The learner will choose QS (.0207) over QI (.00396).

	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	540505	540505
Relativized Probability	**0.00396**	**0.0207**

Table 5.3. Relativize-against-all approach, for both the cues and parsing method. The learner will choose QS.

The relativize-against-potential approach can intuitively be characterized by the question, "How likely is it that a random data point chosen from the set of data points satisfying the preconditions to be unambiguous will actually be an unambiguous data point for the parameter value of interest?" The relativizing set (the set of potential unambiguous data points) will vary in size, depending on whether cues or parsing is used to identify unambiguous data points. Specifically, if the learner uses cues, the relativizing set will vary *across parameter values*. In contrast, if the learner uses parsing, the relativizing set will remain constant across parameter values.

If the learner uses cues to identify unambiguous data, the learner is looking for a combination of structure and stress within a word (e.g. words of 2 syllables that are both stressed for QS). Words that do not match the structural requirement of the cue (e.g. word of 4 syllables for the QS cue) cannot possibly have the correct structure and stress combination to be a cue, since they already lack the correct structure. Thus, these data points are excluded from the set of potential unambiguous data points since they do not obey the necessary structural preconditions. Because of the different structural requirements of the cues for different parameter values, the relativizing set size will vary from cue to cue.

As a concrete example, suppose the data set provides 11213 data points perceived by the learner as unambiguous for QS and 2140 data points perceived by the learner as unambiguous for QI. Suppose also that the potential set of QS cues (words having 2 syllables, etc.) is 85268 while the set of potential QI cues (words of at least 3 syllables, etc.) is 2755. The relativized probability for an unambiguous QS data point is 11213/85268 = 0.132 and the relativized probability for an unambiguous QI data point is 2140/2755 = .777. The learner using the cues method will choose QI (.277) over QS (.132).

	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	2755	85268
Relativized Probability	**0.777**	**0.132**

Table 5.4. Relativize-against-potential approach, for the cues method. The learner will choose QI.

If the learner uses parsing to identify unambiguous data, the set of potential cues consists of all parseable words. The number of parseable words will depend on the currently set parameter values, since some words may not be able to be parsed once certain parameter values are set (e.g. words with syllable-final stress will not be parseable if Em-Right (extrametricality on the rightmost syllable) is set). However, in contrast with the cues method, the size of the relativizing set (the parseable words) will *not* vary from parameter value to parameter value. Thus, all unambiguous data point counts are normalized against the same value, just as in the relativize-against-all approach (though the actual value is less than the entire input set).

As a concrete example, suppose the data set provides 11213 data points perceived by the learner as unambiguous for QS and 2140 data points perceived by the learner as unambiguous for QI. Suppose also that there are *p* parseable words, given the current parameter settings. The relativized probability for an unambiguous QS data point is 11213/*p*, which will be larger than the relativized probability for an unambiguous QI data point, 2140/*p*. The learner using the parsing method will choose QS (11213/*p*) over QI (2140/*p*).

	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	*p*	*p*
Relativized Probability	**2140/*p***	**11213/*p***

Table 5.5. Relativize-against-potential approach, for the parsing method. The learner will choose QS.

5.7.3.3 An Example of Testing a Parameter-Setting Order

In (29), I demonstrate steps (28a-b) for a learner using the parsing method and the relativize-against-all approach, testing a parameter-setting order that begins by setting the quantity sensitivity parameter to QS. Note how the distribution of unambiguous data for a given parameter (such as Extrametricality below) can shift drastically, depending on what parameters are currently set.

(29) Testing an order for the parsing method with the relativize-against-all approach that begins by setting quantity sensitivity

(a) Currently unset parameters: Quantity Sensitivity, Extrametricality, Feet Directionality, Boundedness, Feet Headedness

Quantity Sensitivity		**Extrametricality**	
QI: 0.00398	QS: 0.0205	Em-None: 0.0284	Em-Some: 0.0000259
Feet Directionality		**Boundedness**	
Ft Dir Left: 0.000	Ft Dir Rt: 0.00000925	Unb: 0.00000370	Bounded: 0.00435
Feet Headedness			
Ft Hd Left: 0.00148	Ft Hd Rt: 0.000		

Table 5.6. Unambiguous data distribution from corpus: probability of finding unambiguous data point in input data set, using parsing method and relativize-against-all (probabilities calculated out of 540505 words)

(b) Choose quantity sensitivity to set. QS has a higher probability of finding an unambiguous data point (QS probability is 0.0205, which is greater than QI's probability of 0.00398). Set Quantity Sensitivity to QS.

(c) Currently unset parameters: QS-VC-Light/QS-VC-Heavy, Extrametricality, Feet Directionality, Boundedness, Feet Headedness

QS-VC-Light/QS-VC-Heavy	**Extrametricality**

VC-Light: 0.00265	VC-Heavy: 0.00309	Em-None: 0.0240	Em-Some: 0.0485
Feet Directionality		**Boundedness**	
Ft Dir Left: 0.000	Ft Dir Rt: 0.00000555	Unb: 0.00000370	Bounded: 0.00125
Feet Headedness			
Ft Hd Left: 0.000588	Ft Hd Rt: 0.0000204		

Table 5.7. Unambiguous data distribution from corpus: probability of finding unambiguous data point in input data set, using parsing method and relativize-against-all (probabilities calculated out of 540505 words)

This process then continues for the remaining unset parameters in the system until all parameters are set.

5.8 English Learning Results

5.8.1 Order Constraints as a Metric

If both learning methods yield a set of parameter-setting orders that lead to the correct target values for English, then both solve the logical problem of language learning for the English metrical phonology system. That is, both have at least one parameter-setting order that leads the learner to the target state. If there is more than one viable order, we can then compare the two methods by how well-formed the sets of viable parameter-setting orders are.

First, we can determine if the set for each method can be captured by order constraints at all, whether stipulated or principled. If so, then the set is at least well-formed enough to be described in a more compact representation than explicitly listing all the viable orders in the set. After that, we can then consider the nature of the order constraints that capture each set. A method with a set that can be described by principled constraints will be considered superior to a method with a set that can only be described by constraints that must be explicitly stipulated.

5.8.2 Parameter-setting Orders that Lead to English Target Values

As it turns out, both methods yield a set of parameter-setting orders that will cause a learner to converge on the English values when the relativize-against-all approach is used to calculate the relativized probability of unambiguous data. Both methods thus pass the first hurdle of solving the logical problem of language learning for the English metrical phonology system. Again, this is no mean feat given the interactive nature of the 9 parameters that produce stress contours and the noisiness of the data to the learner.

However, only the parsing method succeeds when the relativize-against-potential approach is used. Because the relativizing set for parsing is constant across parameter values for both relativization approaches, the set of viable orders for parsing is the same for each approach. In contrast, the relativizing set for cues varies across parameter values in the relativize-against-potential approach, and in fact leads to *no* orders being viable to reach the target state for English. Table 5.8 summarizes the behavior of the cues and parsing methods when combined with different approaches to relativizing the probability of the unambiguous data.

	Cues	Parsing
relativize-against-all	**Successful**	**Successful**
relativize-against-potential	*Unsuccessful*	**Successful**

Table 5.8. Comparison of success of different methods of identifying unambiguous data with different approaches to relativizing probability of unambiguous data.

Given that the parsing method always has a viable set of orders, one might believe that parsing is therefore the superior method for identifying unambiguous data. It succeeds no matter what the probability is relativized against because the relativizing set is constant across all parameter values. However, recall that the characterization of the viable set of orders is also important. A set characterized by constraints that are principled is more desirable than a set characterized by constraints that must be stipulated. I shall therefore examine the viable set of orders for both methods and see how they compare.

In (30a) below, I list a sample of the parameter-setting orders for the cues method that allowed the learner to converge on the English values. In (30b), I list a sample of the parameter-setting orders that failed to work. For a complete listing of the orders that were successful, see the Appendix.

(30a) Sample of Cues Method Parameter-Setting Orders that Succeeded

(a) QS, QS-VC-Heavy, B, B-2, Feet Hd Left, Feet Dir Right, Em-Right, B-Syl

(b) QS, B, B-2, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, Em-Right, B-Syl

(c) B, B-2, Feet Dir Right, QS, Feet Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl

(d) Feet Hd Left, Feet Dir Right, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl

(e) Feet Dir Right, QS, Feet Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl

(30b) Sample of Cues Method Parameter-Setting Orders that Failed

(a) Em-Some, Em-Right, Feet Dir Right, QS, Feet Hd Left, B, QS-VC-Heavy, B-2, B-Syl

(b) QS, B, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Bounded-2

(c) Feet Hd Left, Feet Dir Right, B, B-Syl, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right

In addition to a viable set of parameter-setting orders existing for the cues method, this viable set can also be described more succinctly by the order constraints in (31).

(31) Cues Method: Order Constraints

(a) QS-VC-Heavy set before Em-Right
(b) Em-Right set before B-Syl
(c) B-2 set before B-Syl
The rest of the parameters are freely ordered with respect to each other.

In (32a) below, I list a sample of the parameter-setting orders for the parsing method that allowed the learner to converge on the English values. In (32b), I list a sample of the parameter-setting orders that failed to work. For a complete listing of the orders that were successful, see the Appendix.

(32a) Sample of Parsing Method Parameter-Setting Orders that Succeeded

(a) QS, B, Feet Hd Left, QS-VC-Heavy, Feet Dir Right, B-Syl, B-2, Em-Some, Em-Right
(b) B, QS, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, B-Syl, Em-Some, Em-Right, B-2
(c) Feet Hd Left, QS, QS-VC-Heavy, B, Feet Dir Right, En-Some, Em-Right, B-Syl, B-2

(32b) Sample of Parsing Method Parameter-Setting Orders that Failed

(a) Feet Dir Right, QS, Feet Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl
(b) Em-Some, Em-Right, QS, B, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, B-Syl, B-2
(c) QS, QS-VC-Heavy, Feet Hd Left, Feet Dir Right, B, B-Syl, B-2, Em-Some, Em-Right
(d) QS, Feet Dir Right, QS-VC-Heavy, Feet Hd Left, B, B-Syl, B-2, Em-Some, Em-Right
(e) B, Feet Dir Right, QS, QS-VC-Heavy, Feet Hd Left, B-Syl, B-2, Em-Some, Em-Right

In addition to a viable set of parameter-setting orders existing for the parsing method, the viable set can also be described more succinctly by dividing the parameters into three groups. The parameters within each group are freely ordered with respect to each other (33).

(33) Parsing Method: Order Constraints as Freely-Ordered Groups

(a) Group 1: QS, Ft Hd Left, B
(b) Group 2: Ft Dir Right, QS-VC-Heavy
(c) Group 3: Em-Some, Em-Right, B-2, B-Syl

At first glance, the order set for the cues method appears to be less constrained than the order set for the parsing method. However, the true criterion of merit is to compare how easily each of the constraints can be derived from other properties of the learning system.

5.8.3 Deriving Constraints

There are several ways I could think of to derive constraints from properties of the learning system: data saliency, data quantity, and default values.[52] I describe each of these in turn.

I begin with the saliency of the data. Data that are better signals might be noticed and used more easily by the learner than data that aren't. This is true no matter what the domain. In the domain of metrical phonology, it has been suggested that the unexpected presence of stress is more informative than the unexpected absence of stress (Bill Idsardi, *personal communication*). The presence of stress is a stronger logical signal since there are many factors that could cause the absence of stress if the stress system is unknown, e.g. stress deletion under clash, conflation of secondary stresses, and segmental rules such as vowel devoicing (Halle & Idsardi, 1995; among others). The presence of stress, however can pinpoint a parametric cause (or a lexically pre-existing stress that has to be stored explicitly in the system anyway (Halle & Idsardi, 1995)). Moreover, the presence of stress may be a stronger acoustic signal, since a stressed syllable is more prominent than an unstressed syllable. Stressed syllables might therefore be more readily attended to by the learner.

There is also morphological evidence that the presence of stress is psychologically more salient. Morphological rules exist that restrict affix attachment to words with stress on the edge syllable (-al for final stress words: remove + al = removal), but I am currently unaware of any morphological rules that exist that restrict affix attachment to words *without* stress on the appropriate syllable. This suggests some psychological priority for paying attention to stressed syllables over their unstressed counterparts. Given the informational asymmetry between the presence and absence of stress, we might expect parameters that rely on the learner noticing the absence of stress to be deprioritized. Extrametricality (Em-Some, Em-Right, Em-Left) is just such a parameter; thus, we might expect it to be set later than other parameters.

Secondly, the quantity of data available to the learner could also affect parameter-setting order. Again, this will be true irrespective of the domain. Parameters with more unambiguous data available are likely to be set before parameters with less, simply because there is more data for the learner to use for updating.

Thirdly, if the learner is using a default value, we can dispense with constraints for that value if it is the correct adult value since it is already set by default. Again, this will be true irrespective of the domain. The logic behind this is that a constraint of the form "Parameter value A1 must be set before parameter value B1" results from either (a) A1 not being able to be set correctly if B1 is set first (i.e. the unambiguous data distribution favors A2 after B1 is set) or (b) B1 not being able to be set correctly until A1 is set (i.e. the unambiguous data distribution favors B2 until A1 is set). Depending on which it is, this problem can be side-stepped if either

[52] There may in fact be more as well. These three come to mind as being fairly general properties of the learning system.

(a) A1 is the default for A or (b) B1 is the default for B, respectively, since the correct parameter value is already set. The constraint "Parameter value A1 must be set before parameter value B1" is then unnecessary.

I note that using defaults only applies to the cues method since the instantiation of the parsing method used here must use all available values to conduct a find-all-parses analysis. One might argue that the parsing method could in fact be instantiated with a default system under a different implementation. However, this has an inherent problem. Specifically, the only values available to the parser would be the default values. Thus, only parses using the default values would be considered by the learner initially. This is fine if the adult values for the system are the default values. But, suppose they are not. How will the learner recognize unambiguous data for the non-default values, a problem noted by Valian (1990)? The parser, by definition, can only use data it can parse. The non-default values are not in its set of available values, and so it will not be able to parse data that can only be parsed with those values. In short, the parsing method cannot comprehend data that are unambiguous for the non-default values since it cannot parse such data with the default values. This is in sharp contrast to the cues method, which can still recognize unambiguous data for the marked values even while the default values are in place.

I will now examine which constraints for the cues and parsing methods can be accounted for using these three explanations: data saliency, data quantity, and default values.

5.8.3.1 Cues Method with Relative-Against-All: Accounting for Constraints

The first constraint (31a) was that QS-VC-Heavy must be set before Em-Right. We can derive this via data saliency, and argue that noticing the absence of stress for extrametricality is more difficult than noticing the presence of stress in the pattern for QS-VC-Heavy.

The second constraint (31b) was that Em-Right must be set before Bounded-Syl. (This is due to Bounded-Mor being favored until Em-Right is set.) When we examine that unambiguous data distribution, it turns out that Em-Right has at least 20 times as much data as Bounded-Syl (and so, the learner is 20 times more likely to find an Em-Right cue) at any given point in time. Thus, this constraint could be derived from data quantity. Also, I noted in section 5.4.4 that the cues learner could use Bounded-Syl as a default value once the more general Bounded value is set. If this is the case, then Bounded-Syl will already be set and this constraint disappears from the use of default values.

The third constraint (31c) was that Bounded-2 must be set before Bounded-Syl. (Bounded-Mor is favored until Bounded-2 is set.) Unfortunately, the unambiguous data distribution favors Bounded-Syl over Bounded-2 initially so we cannot directly derive this constraint from data quantity. However, there is a partial ordering with Em-Right which can be useful. Specifically, once Em-Right is set, a Bounded-2 cue is at least 4 times as likely to be found as a Bounded-Syl cue at any given point and would then be set first. So, once Em-Right is set, this constraint can be derived from data quantity. However, this requires Em-Right to be set before Bounded-2.

Fortunately, an Em-Right cue is about 270 times more probable than a Bounded-2 cue, so Em-Right could easily be set first. Thus, this constraint could be derived from data quantity: set Em-Right, then Bounded-2, and then Bounded-Syl. Also, we could rely on default values again to cause this constraint to disappear: Bounded-Syl is the default value once the more general Bounded value is set.

What is striking here is that *all* of the cues method order constraints are derivable from other properties of the learning system (either the learner's learning preferences or the available data). They do not need to be explicitly stated or available to the learner as pre-specified knowledge. This makes these constraints highly attractive.

5.8.3.2 Parsing Method with Relative-Against-All/Potential: Accounting for Constraints

The parsing method's constraints, however, are not so easily derived. Recall that the parsing method learner must constrain parameter-setting to three parameter groups that are ordered with respect to each other (33) – all the ones in the first group (QS, Feet Hd Left, Bounded) must be set before all the ones in the second group (Ft Dir Right, QS-VC-Heavy), and all the ones in the second group must be set before all the ones in the third group (Em-Some, Em-Right, Bounded-2, Bounded-Syl). Since the parsing method learner in this model cannot use default values, the constraints can be derived only from the properties of data saliency and data quantity.

I note that even supposing the parsing method *could* somehow use default values, these constraints still cannot all be derived. The only constraint that default values could account for is Bounded-Syl in the third grouping: Bounded-Syl is the default value, and so would already be set. There is no need for it to be set after other parameters. No other constraints could be accounted for by default values since the adult values are the non-default values (QS, QS-VC-Heavy, Bounded, Em-Some).

Still we can ask how much can be accounted for by the remaining two properties. Data saliency will explain why Em-Some and Em-Right are in the last group: noticing the absence of stress puts these parameters later in the learning path (group 3). This leaves data quantity to account for all the rest. Unfortunately, data quantity will not separate the remaining parameters into the three necessary groups. A parsing method learner would need to have these groups explicitly built in as prior knowledge, which makes these constraints less attractive than their cues method counterparts. The ability to derive all of the relevant order constraints thus seems to favor the cues method, when used with the relativize-against-all approach.

The success of the cues and parsing methods are compared below in Table 5.9. As we saw previously, the parsing method seems more flexible because it succeeds no matter what relativization approach is used. The cues method, however, has a set of order constraints that can be derived from properties of the learning system when this method does actually succeed.

	Reaches Target State	All Order Constraints Derivable
Cues + Relative-Against-All	Yes	Yes
Parsing + Relativize-Against-All	Yes	No
Parsing + Relativize-Against-Potential	Yes	No
Cues + Relativize-Against-Potential	No	*N/A*

Table 5.9. Comparing the performance of the cues and parsing methods, when used with different relativization approaches. The optimal combination for this case seems to be the cues method with the relative-against-all approach, since it both reaches the target state and has derivable order constraints.

5.9 Discussion

5.9.1 Cues: Why Better Constraints on Parameter-Setting Order?

As we saw in the previous section, the cues method results in a set of parameter-setting orders that can be captured by constraints that are independently derivable and few in number. This is not true for the order constraints that capture the parsing method's set: that set is far more restricted, and requires a larger number of constraints, most of which must be stipulated. I speculate that this has to do with the nature of the data that a cues method learner uses.

Cues themselves are small pieces of highly informative surface structure, such as 2 syllable words with 2 stresses (QS, QS-VC-Heavy, Bounded-Mor) or the leftmost syllables in a word with stress in a certain pattern (Ft Dir Rt, Ft Hd Left). Crucially, the learner doesn't have to understand the entire data point to identify a cue in the data point. In fact, the data point can be in conflict with values that are already set but *still* contain cues for currently unset values.

For example, a 2 syllable word with 2 stresses is in conflict with Em-Right since it has stress on the rightmost syllable, but is still useful as a cue for QS. This gives a cues method learner more flexibility than a parsing method learner has, since the cues learner can make use of the non-problematic portions of data points instead of having to disregard these portions along with the entire data point.

For the parsing method learner, if a data point can't be parsed (because the learner doesn't understand the entire data point or the data point is in conflict with currently set values), the data point can't be used at all. Note that this problem persists even when using other less resource-intensive parsing strategies (Sakas & Fodor, 2001) since those strategies consider cases where multiple parses can describe the data point, but not cases where *no* parses describe the complete data point. Unless the parsing method can retrieve information from only a subpart of the data point, the problem that plagues the parsing method here will persist. The noisiness of the English metrical phonology data set greatly penalizes the parsing method learner, which is reflected in the greater quantity of order constraints required to capture the more restricted set of viable parameter-setting orders.

However, the flexibility of cues is not without its drawbacks – a cues-learner can be led irrecoverably astray in some cases as we saw previously. When the cues method is combined with the relativize-against-potential approach, certain values that are not in the English target state persist no matter what other values are set. For instance, because the relativizing set of QI unambiguous data is significantly smaller than that of the QS unambiguous data (QI: 2755, QS: 85268), a cues learner using the relativize-against-potential approach consistently awards a higher probability to the QI unambiguous data. No other parameter settings will influence the potential QI set because the QI cue does not interact with any other parameter value (e.g. the way QS cues do with Extrametricality), and so it will *always* be significantly smaller than the potential QS set. Because no other parameter settings affect the cue for the QI value, the relativizing QI set can never be altered. Unfortunately for a learner of English, having such a small relativizing QI set will cause the learner to favor the QI unambiguous data over the QS unambiguous data. Since QS is the correct value for English, no viable parameter-setting orders exist for cues when using the relativize-against-potential approach. Thus, we see that the flexibility the cues method has can be both a strength and a weakness, depending on what other learning strategies the learner adopts. Nonetheless, it is this flexibility which yields a more concise representation of knowledge necessary for acquisition success (the order constraints) when the method does, in fact, succeed.

5.9.2 Relativization

I examined two different approaches a learner might adopt for relativizing the probability of an unambiguous data point for a given parameter value: relativize-against-all and relativize-against-potential. While we had no a priori reason for assuming one approach was superior to the other, we may wish to use the results obtained here to support the relativize-against-all approach. Specifically, in order to reach the target state and have a set of viable orders that can be described by a small set of principled order constraints, a learner must use the cues method coupled with the relative-against-all approach. Thus, the learning procedure relativizes the probability of an unambiguous data point against the entire set of input seen so far. This is in contrast to a learning procedure that keeps track of the quantity of potential unambiguous data points, and relativizes the probability of an unambiguous data point against that set (which will vary across parameter values for cues). Because the learner does not need to keep track of the set of potential unambiguous data points for each parameter value, the relativize-against-all approach is likely less resource-intensive to implement as well. This is a desirable quality for a psychologically plausible learning strategy.

5.9.3 Cues and Parsing: A Viable Combination?

Cues and parsing have a complementary array of strengths and weaknesses as methods for identifying unambiguous data. From the case study examined here, we have seen an additional strength and weakness for both cues and parsing. Cues give us a principled set of order constraints, but aren't robust across different strategies of relativization. The opposite is true for parsing: we find robustness across different relativization approaches, but a set of order constraints that must be mostly stipulated. We also examined additional weaknesses in section 5.2 for both methods. Cues are knowledge the learner must have already available; parsing can only use the entire data point, rather than just a subpart.

A very interesting question is if there is a way to combine these two methods to capitalize on their complementary strengths and mitigate their complementary weaknesses. I speculate now on how this might be accomplished. Cues themselves are small pieces of highly informative surface structure that are usually smaller than the entire data point. Given this, perhaps a learner might derive cues from a limited kind of parsing (perhaps limited by time and mental resources available). Such a limited parsing method could be biased to use subparts of a data point rather than trying to assign full parses to the entire data point.

For example, suppose a learner with no values set hears a sequence of syllables in the speech stream and realizes that two consecutive syllables are the beginning of a new word.[53] The learner then tries to analyze these two syllables alone. Suppose the first of these two syllables is stressed and contains a long vowel (VV) while the second is unstressed and contains a short vowel with a coda (VC).

(34) speech stream, with two syllables of new word (signaled by #): …# VV VC…

The learner then tries to parse these two syllables with any parameter values that can be applied, given that only the beginning of the word is known. (It is possible that these two syllable comprise the entire word, but the learner is unaware of this.) The learner then tries to parse this sequence of syllables with all *applicable* parameter values – i.e., values that can apply to the front subpart of a word alone. The set of applicable values would be Quantity-Insensitive, Quantity-Sensitive, Extrametricality-Some [Left], Unbounded, Bounded, Bounded-Syllabic, Bounded-Moraic, Feet Headed Left, Feet Headed Right, and Feet Directionality Left. Em-None and Em-Right are not applicable since the right edge of the word is unknown, so nothing can be observed about the final syllable. Feet Dir Right is also not applicable for similar reasons: the learner cannot construct feet starting from the right edge since the right edge is unknown.

[53] Note that there may be some interleaving of learning the metrical phonology system and learning to segment words successfully. Learners early on have a sense of the basic rhythmic properties of their language (Mehler et al. 1988, Nazzi et al., 2000) – for instance, trochaic (first syllable stressed) or iambic (second syllable stressed) as stereotypical (Jusczyk et al. 1993). They may then use this highly probable rhythmic pattern to segment syllables in the speech stream into words (Jusczyk et al, 1999; Houston et al., 2000; Houston et al., 2004). Sometimes, this will result in mis-segmentation: *ba na na* becomes segmented as simply *na na* in English. This could then lead to misanalysis in the more complex metrical phonology domain, since the "word" being analyzed isn't actually the word in the target language (analysis of "nana" instead of "banana"). If the more elaborate metrical phonology system examined here is learned early enough that correct word segmentation isn't regularly successful, this could be another factor that determines learners' success. In effect, they are applying an additional filter to the available input and only perceiving words that match the basic rhythmic bias they have acquired already. Thanks to the CUNY Supper Club for very useful discussion of this point.

(35) Viable Partial Parses for the syllable sequence #<u>VV</u> VC…

(a) (QI, **Ft Dir Left**, Unb, **Ft Hd Left**)
(b) (QI, **Ft Dir Left**, B, B-Syl, B-2, **Ft Hd Left**)
(c) (QI, **Ft Dir Left**, B, B-Syl, B-3, **Ft Hd Left**)
(d) (QS, QS-VC-Light, **Ft Dir Left**, Unb, **Ft Hd Left**)
(e) (QS, QS-VC-Light, **Ft Dir Left**, B, B-Syl, B-2, **Ft Hd Left**)
(f) (QS, QS-VC-Light, **Ft Dir Left**, B, B-Syl, B-3, **Ft Hd Left**)
(g) (QS, QS-VC-Light, **Ft Dir Left**, B, B-Mor, B-3, **Ft Hd Left**)

Of all the available values, only Feet Headed Left and Feet Directionality Left are used by all parses of this two syllable sequence. Because Feet Directionality Right was not applicable, the learner will not conclude anything about Feet Directionality. Similar reasons preclude the learner from using this data point to signal Extrametricality – the full range of values for that parameter was not applicable: even though the learner would perhaps be able to rule *out* Extrametricality-Left, there is no definitive distinction between Em-None, Em-Some, or Em-Right. However, all the values for Feet Headedness *were* applicable: both Feet Headed Left and Feet Headed Right. Since Feet Headed Left was required for all parses, the learner would perceive this two syllable sequence as unambiguous for Feet Headed Left.

In this way, the learner would be deriving cues from a limited form of parsing that operates over subparts of data points. Note that if the learner derives cues from the implementation of parsing used here, the learner loses the ability to use default values since default values are not compatible with this instantiation of parsing. A learner cannot unlearn default values if the learner only ever uses default values to parse data; data indicating the non-default values are unparseable and therefore cannot be learned from (Valian, 1990). However, it may be possible to sidestep this problem with a probabilistic parser that favors default values and probabilistically uses them for parsing. Then, the learner would still be able to parse (a portion of) the unambiguous data encountered for the non-default value, if the adult system used the non-default value.

Still, we also lose the benefit from parsing that allows probability relativization to be constant across parameter values. The relativize-against-potential approach would have a relativizing set consisting only of the data which that value could possibly have parsed. The example we described above would be included in the relativizing set for Feet Headed Right (since Feet Headed Right was applicable), but not in the relativizing set for Feet Directionality Right (since Feet Directionality Right wasn't applicable).

However, it is possible that using limited parsing to derive cues gains some of benefits associated with using cues in the first place. In particular, operating over subparts of a data point is what I believed allowed the cues method to have fewer order constraints. It's possible that using cues derived from limited parsing would also produce a set of viable orders that can be characterized by fewer constraints. So, I posit that a learner using cues derived from limited parsing would potentially have the desired behavior combining the strengths of parsing and cues: less necessarily innate knowledge and fewer order constraints. This prediction, of course, remains to be explored.

Also, a limited parsing method would likely result in partial analyses that are more heuristic than exact, possibly at the expense of more false positives and false negatives. Though this may seem to be undesirable, such behavior may be good from the perspective of language change since certain language changes require *imperfect* learning, as we saw in the previous chapter. If data the learner considers unambiguous are keyed more to the surface form and are less well-connected to the abstract grammatical parameters, then it is easier for slippage to occur over time.

As a specific example, recall from the previous chapter that the change in Old English from Object-Verb order to Verb-Object order has been argued to be the result of imperfect learning in just this way (Lightfoot, 1991). Learners use cues (or parsing over a limited set of parameters) to find data they perceive as unambiguous, though this data may actually be ambiguous if parsed more fully. This allows the learners to converge on a slightly different probability distribution than the adults of the population have. Specifically for Old English, the system is a probability distribution between Object-Verb and Verb-Object order. The learners end up with a final probability that is marginally different from the probability of the rest of the population. Over time, these small "slips" lead to language change in the population. Importantly for communication purposes, the slips are, as mentioned, *small*. Cues derived from limited parsing would potentially allow learning to be successful enough to achieve the desired target state in most cases, but not so successful that small changes are impossible.

5.9.4 Future Directions

There are several immediate ways to build upon the findings concerning (a) other instantiations of the unambiguous data filter, (b) the sufficiency of the unambiguous data filter for other languages, (c) the necessity of the unambiguous data filter for learning metrical phonology, (d) experimentally testable predictions made by the unambiguous data filter, and (e) distinguishing systematic exceptions from noise in order to form irregular sub-systems given the available data.

The previous section described a potential combination of the methods for identifying unambiguous data that would retain the strengths of both the methods examined, cues and parsing. This combination strategy's ability to actually converge on the English system should be examined, as well as any constraints required for its success. When I examined the cues and parsing methods separately, each required different constraints for acquisition success on the English dataset: cues required a particular assumption about how the learner relativizes the probability of unambiguous data, while parsing required order constraints that would need to already be available to the learner. The combination strategy might require constraints of both kinds (probability relativization and prior knowledge of parameter-setting order), one kind, or neither kind.

From the perspective of the logical problem of language learning, future work could also test the cues, parsing, and limited parsing methods on other languages for which we have sufficient corpora of child-directed speech. These methods can also be investigated in other domains besides metrical phonology.

The necessity of the unambiguous data filter also can be examined for this case study. As in the previous chapter's future directions, there are various ways to relax the unambiguous data filter and have the learner use ambiguous data. For instance, the learner could weight ambiguous data points such that they're not as influential as unambiguous data (again, as done in chapter 3 for learning anaphoric *one*). For the parsing method, the learner might adopt a probabilistic weighting of ambiguous data based on the percentage of successful parses that share a certain value. As an example, suppose 4 of 5 successful parses for a data point require Extrametricality-None, while 1 requires Extrametricality-Some. The learner might then give 80% credit to Extrametricality-None and 20% credit to Extrametricality-Some.

The learner might also adopt an ambiguous data strategy that probabilistically chooses one parameter value for each parameter to parse the data point. Successful parses reward all the parameter values used while unsuccessful parses punish all the parameter values used, as instantiated in the Naïve Parameter Learning model of Yang (2002). As an example, suppose the learner encounters an ambiguous data point and only has Bounded-2 vs. Bounded-3 and Extrametricality-Right vs. Extrametricality-Left remaining to be set. Suppose also that Bounded-2 is favored over Bounded-3, with associated probabilities of .8 (B-2) and .2 (B-3), and Extrametricality-Right is similarly favored over Extrametricality-Left, .8 (Em-Right) to .2 (Em-Left). Then, the learner chooses one of the four combinations of parameter values to parse the data point with, based on their associated combined probability: (a) B-2, Em-Right (.8*.8 = .64), (b) B-3, Em-Right (.2*.8 = .16), (c) B-2, Em-Left (.8*.2 = .16), (d) B-3, Em-Left (.2*.2 = .04). If the combination of values yields a successful parse of the data point, all values are rewarded; if the parse fails, all values are punished. This learning strategy is implicitly driven by the unambiguous data in the input since unambiguous data for one parameter value (e.g. Em-Right) will be unparseable by the opposing value (e.g. Em-Left), and so punish the opposing value (e.g. Em-Left). However, this strategy does not explicitly seek unambiguous data nor does it restrict the learner to use only unambiguous data, allowing it to avoid the sparse data problem that could potentially plague an unambiguous data learner.

In addition, the unambiguous data filter explored here makes testable predictions about which parameters should be set first in a given language, based on the order constraints required for acquisition success for either method of identifying unambiguous data. For instance, both cues and parsing would predict that a learner should set Quantity Sensitivity before Extrametricality. These predictions can be tested with both modeling (by using realistic estimates of the quantity of data children are exposed to) and standard experimental techniques for infants such as head-turn preference (Jusczyk & Aslin (1995)).

For the modeling extension, we can also investigate whether the necessary order constraints leading to the correct English target state (e.g. a cues learner setting Extrametricality before Bounded-Syllable) can emerge with a high probability simply from the distributions of data available to children or if instead data saliency explanations and/or default values are required. If default values are required, this suggests a prior probability distribution that strongly favors the default value. Moreover, the situation where the learner has a strong bias for one value over another may be analogized to second-language learning: the adult has a very strong initial bias for the native language values. Exploring the behaviors produced from strong initial biases as well as ways to recover from these strong initial biases can have implications for second language learning.

Finally, the current learning model can be extended to search for systematic exceptions in the data in order to form irregular sub-classes. Exceptions (and errors) would be recognized once the learner has some of the system known. For instance, if the learner has determined the English system is Quantity Sensitive, an exceptional data point would be unambiguous for Quantity Insensitive. So, the learner can start recognizing exceptions even before the entire regular metrical system is learned. The learner might then be able to invoke a rule competition model, similar to Yang (2002)'s implementation for forming irregular past tense classes, in order to group irregular metrical phonology data points together into sub-classes. Systematic exceptions would be recognized as distinct from noise (or singular exceptions that should be memorized) based on the frequency of the words – and importantly, the different words – that are exceptional in that way. This again draws from Yang's (2002) implementation of forming irregular classes for the English past tense.

As an example, suppose the learner has decided the main system is Quantity Sensitive. However, the learner then keeps encountering data points that are incompatible with that parameter value: *ponytail*, *ladybug*, *jellybean,* etc. If these examples are frequent enough, the learner might hypothesize that there is an irregular class of words where the second syllable with the long vowel /i/ ('*ny*', /ni/; '*dy*', /di/; '*lly*', /li/) is destressed (even though /i/ is a long vowel and should receive stress given the regular system). If the learner is at a stage in learning where meaning is associated with words, then the irregular class might (additionally) be defined over something like compound words.

Importantly, the learner would need to recognize the exceptional data points as distinct from the main system being learned, but regular enough to warrant positing systematicity for them. To recognize the exceptional data points, the learner must already have some of the parameters for the main system set. The learner would thus benefit from the "preset" parameters of the main system in order to recognize and extract systematicity in any irregular sub-systems that might exist.

5.10 Summary

In this chapter, I have investigated the feasibility of using an unambiguous data filter on the learner's intake for metrical phonology, a complex system with multiple interacting parameters. I have shown that an unambiguous data filter can indeed allow a learner to converge on the correct set of adult values for English metrical phonology, which is a noisy system containing unambiguous data for the incorrect values as well as for the correct values. The learner is successful whether the unambiguous data filter is implemented by using the domain-specific representation of cues (Dresher, 1999; Lightfoot, 1999) or the domain-specific learning procedure of parsing (Fodor, 1998b,1998c; Sakas & Fodor, 2001).

Nonetheless, there are differences between the two methods in terms of what must be explicitly stipulated and what can be derived from the learning system. In addition, the two methods differ on their flexibility across different approaches of relativizing probability.

The parsing method does not need to stipulate additional information to identify unambiguous data, since the domain-specific procedure of assigning structure to a data point is already employed for language comprehension. Moreover, the parsing method succeeds no matter which probability relativization approach is used to analyze the data. Yet, the inability of the parsing method to use default values and make use of subparts of a data point force it to have a more restricted set of viable parameter-setting orders. This in turn leads to order constraints that must be stipulated in the case examined here.

The cues method, on the other hand, *can* use default values and glean information from data point subparts, which allows the set of viable parameter-setting orders to be far less restricted in the case examined here. However, a cues learner can only succeed when the unambiguous data are relativized against the entire input, rendering this method less flexible than the parsing method. Moreover, the original formulation of cues requires us to stipulate the domain-specific knowledge of cues in order to identify unambiguous data.

I have speculated a way of combining both methods: deriving cues from a limited form of parsing that allows parsing over subparts of a data point. The limited parsing method would thus possess two advantageous properties: (1) minimal knowledge is stipulated to identify unambiguous data and (2) more heuristic identification of unambiguous data that could lead to fewer order constraints. It is uncertain, however, if the limited parsing would succeed across different probability relativizations, since the set of potential unambiguous data would vary across parameter values, as it does for the cues implementation examined here. This remains to be explored.

The results obtained here suggest that an unambiguous data filter can lead to the correct learning results in complex domains. The crucial aspect of such a filter is that data are unambiguous *relative* to the learner's perspective, and the learner has incomplete knowledge of the full adult grammar during the learning process. Thus, data that appear unambiguous at an earlier time point may be viewed as ambiguous later when more information has been obtained, and vice versa. Contrary to severely handicapping the learner, such heuristic, inexact definitions of unambiguous data seem to allow the learner the flexibility to triumph in a noisy system. Given that the linguistic environment is often quite noisy, learners may benefit from treating data that conform to their semi-informed definition of unambiguous data as though they were truly unambiguous data – and therefore, fully informative for learning. In this way, a learner can feasibly implement an unambiguous data filter while avoiding the sparse data problem in realistic language learning cases.

Chapter 6: Learning By Filtering

In the case studies presented in this dissertation, I have explicitly investigated one component of the learning theory mechanism: the definition of the data intake. In each case, filtering the data intake has had enormous effects on the output of learning, separating learning failure from learning success. These case studies suggest that, perhaps contrary to intuition, using all the available data for learning isn't what real human learners do. Instead, young children can succeed by using a select subset of data from which they view as more informative and from which it is in some sense easier for them to extract the correct linguistic systematicity. For anaphoric *one*, learners succeed by heeding only the data that is informative about which N' to choose when there is more than one N' antecedent available. For word order properties such as Object-Verb or Verb-Object order, learners succeed by using degree-0 data that they perceive as unambiguous. For the English metrical phonology system, learners again succeed by using data perceived as unambiguous. Data intake restriction is key: using fewer data points that are cleaner is superior to using many data points that are noisy representations of the underlying linguistic system.

The division of the learning theory into distinct components allows us to combine components of different types together: domain-specific and domain-general, discrete and probabilistic. Moreover, this framework is a tool that can be applied to many learning problems with different hypothesis spaces that combine information across domains. In this dissertation, I have applied it to subset-superset hypotheses in the syntax-semantics interface, probabilistic distributions between hypotheses in syntax, and multiple interacting hypotheses in metrical phonology. In addition, the distinct components of the framework can be investigated separately, as I do here for data intake filtering. For this investigation, computational modeling has been a very valuable tool, since it allows precise control over the learner's data intake in a way that is difficult to achieve with traditional experimental techniques.

In sum, this dissertation represents the first steps towards a theory of the mechanism of language learning. I have answered the more specific questions set out for each case study. Yet, this has opened the way for still more questions. Future work, especially computational modeling work, will hopefully continue to draw on both theoretical and experimental linguistic data to explore how language learning can succeed in the noisy environment that surrounds young learners.

Appendix

This is the list the complete set of parameter-setting orders for each method and relativization approach that allowed the learner to converge on the English metrical phonology parameter values. From these sets, the order constraints described in section 5.7 were derived.

(A1) Viable Parameter-Setting Orders for the Cues Method, Relativize-Against-All

(QS, QS-VC-Heavy, B, B-2, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, B-2, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, B-2, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, B-2, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left, Ft Dir Rt)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt, Ft Hd Left)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, Ft Hd Left, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, Ft Dir Rt, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left, Ft Dir Rt)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt, Ft Hd Left)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Hd Left, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Dir Rt, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Hd Left, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Dir Rt, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-2, B-Syl)

(QS, QS-VC-Heavy, Ft Hd Left, B, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, B, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, B, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Ft Hd Left, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, B, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, B, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Ft Hd Left, B, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, B, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Ft Dir Rt, B, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, B, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Ft Dir Rt, B, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, B, B-2, Em-Some, Em-Right, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B, Ft Hd Left, B-2, B-Syl)
(QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Hd Left, Ft Dir Rt)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Dir Rt, Ft Hd Left)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Hd Left, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Hd Left, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Dir Rt, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Dir Rt, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Hd Left, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Hd Left, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Hd Left, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Dir Rt, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Dir Rt, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Dir Rt, Ft Hd Left, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B, B-2, Ft Dir Rt, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B, B-2, B-Syl, Ft Dir Rt)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B, Ft Dir Rt, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B, B-2, Ft Hd Left, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B, B-2, B-Syl, Ft Hd Left)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B, Ft Hd Left, B-2, B-Syl)
(QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B, B-2, B-Syl)
(QS, B, QS-VC-Heavy, B-2, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-Syl)

(QS, B, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, B-2, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, B, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left, Ft Dir Rt)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt, Ft Hd Left)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left, Ft Dir Rt)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt, Ft Hd Left)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, Ft Dir Rt, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl, Ft Dir Rt)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-2, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, Ft Hd Left, B-Syl)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl, Ft Hd Left)
(QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-2, B-Syl)
(QS, B, B-2, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, B, B-2, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, B, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, B, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, B, B-2, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, B, B-2, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, B, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, B, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, B, Ft Hd Left, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, B, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)

(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, B, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, B, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, B, Ft Hd Left, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, Ft Hd Left, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, B, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, B, Ft Dir Rt, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, B, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, B, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, Ft Dir Rt, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, Ft Hd Left, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Dir Rt, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Dir Rt)
(QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Dir Rt, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B, B-2, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(QS, Ft Hd Left, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(QS, Ft Hd Left, B, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, B, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)

(QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, Ft Hd Left, Ft Dir Rt, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, Ft Dir Rt, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, Ft Dir Rt, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, QS-VC-Heavy, B, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, B, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Hd Left, B-2, B-Syl)
(QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B, B-2, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, B, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(QS, Ft Dir Rt, B, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, B, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(QS, Ft Dir Rt, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(QS, Ft Dir Rt, B, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, B, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, B, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Dir Rt, Ft Hd Left, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, QS, QS-VC-Heavy, B-2, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left, Ft Dir Rt)
(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt, Ft Hd Left)
(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl, Ft Dir Rt)

(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, QS, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(B, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left, Ft Dir Rt)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt, Ft Hd Left)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl, Ft Dir Rt)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, Ft Dir Rt, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl, Ft Dir Rt)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, Ft Dir Rt, B-2, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, Ft Hd Left, B-Syl)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl, Ft Hd Left)
(B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, Ft Hd Left, B-2, B-Syl)
(B, QS, B-2, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, B-2, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, B-2, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, QS, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, QS, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, QS, Ft Hd Left, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(B, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(B, QS, Ft Hd Left, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Hd Left, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)

(B, QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(B, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(B, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(B, QS, Ft Dir Rt, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Dir Rt, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, QS, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Dir Rt, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, QS, Ft Dir Rt, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, QS, Ft Dir Rt, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, B-2, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, B-2, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, B-2, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, B-2, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, B-2, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, B-2, Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, B-2, Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, B-2, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, B-2, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, B-2, Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(B, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(B, Ft Hd Left, QS, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Hd Left, QS, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, B-2, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, B-2, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(B, Ft Hd Left, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(B, Ft Hd Left, B-2, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Hd Left, Ft Dir Rt, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, Ft Dir Rt, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, QS, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, Ft Dir Rt, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)

(B, Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(B, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(B, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(B, Ft Dir Rt, QS, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, QS, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, Ft Dir Rt, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Dir Rt, QS, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, B-2, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, B-2, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(B, Ft Dir Rt, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(B, Ft Dir Rt, B-2, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Dir Rt, Ft Hd Left, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(B, Ft Dir Rt, Ft Hd Left, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(Ft Hd Left, QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Dir Rt, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Dir Rt)
(Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Dir Rt, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B, B-2, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(Ft Hd Left, QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(Ft Hd Left, QS, B, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, B, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, B, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)
(Ft Hd Left, QS, Ft Dir Rt, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Hd Left, QS, Ft Dir Rt, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, Ft Dir Rt, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, B-2, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(Ft Hd Left, B, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Dir Rt, B-Syl)
(Ft Hd Left, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Dir Rt)
(Ft Hd Left, B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-2, B-Syl)
(Ft Hd Left, B, QS, B-2, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, B, QS, Ft Dir Rt, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, B-2, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, B-2, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Dir Rt, B-Syl)
(Ft Hd Left, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Dir Rt)
(Ft Hd Left, B, B-2, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, Ft Dir Rt, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, B, Ft Dir Rt, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, B, Ft Dir Rt, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, Ft Dir Rt, QS, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, Ft Dir Rt, B, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, Ft Dir Rt, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)
(Ft Dir Rt, QS, QS-VC-Heavy, B, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)
(Ft Dir Rt, QS, QS-VC-Heavy, B, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, B, B-2, Em-Some, Em-Right, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, B, Em-Some, Em-Right, B-2, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B, B-2, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, Ft Hd Left, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl, Ft Hd Left)
(Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, B, Ft Hd Left, B-2, B-Syl)
(Ft Dir Rt, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B, B-2, B-Syl)
(Ft Dir Rt, QS, B, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)
(Ft Dir Rt, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)
(Ft Dir Rt, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)

(Ft Dir Rt, QS, B, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, B, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)

(Ft Dir Rt, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)

(Ft Dir Rt, QS, B, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)

(Ft Dir Rt, QS, B, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, B, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)

(Ft Dir Rt, QS, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)

(Ft Dir Rt, QS, B, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, B, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, QS, B, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, QS, Ft Hd Left, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, B-2, Ft Hd Left, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, Ft Hd Left, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl, Ft Hd Left)

(Ft Dir Rt, B, QS, QS-VC-Heavy, Ft Hd Left, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, Ft Hd Left, B-Syl)

(Ft Dir Rt, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl, Ft Hd Left)

(Ft Dir Rt, B, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-2, B-Syl)

(Ft Dir Rt, B, QS, B-2, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, B-2, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)

(Ft Dir Rt, B, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)

(Ft Dir Rt, B, QS, Ft Hd Left, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, B, QS, Ft Hd Left, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, B-2, QS, Ft Hd Left, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, B-2, QS, QS-VC-Heavy, Ft Hd Left, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, Ft Hd Left, B-Syl)

(Ft Dir Rt, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, Ft Hd Left)

(Ft Dir Rt, B, B-2, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, Ft Hd Left, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, B, Ft Hd Left, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, B, Ft Hd Left, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, B, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, B, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, QS-VC-Heavy, Em-Some, Em-Right, B, B-2, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, B, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, B, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, Ft Hd Left, QS, B, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, B, QS, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, B, QS, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)

(Ft Dir Rt, Ft Hd Left, B, QS, B-2, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(Ft Dir Rt, Ft Hd Left, B, B-2, QS, QS-VC-Heavy, Em-Some, Em-Right, B-Syl)

(A2) Viable Parameter-Setting Orders for the Parsing Method, Relativize-Against-All

(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(QS, B, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-Syl, B-2, Em-Some, Em-Right)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, B-Syl, Em-Some, Em-Right)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-Syl, Em-Some, Em-Right, B-2)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, B-2)
(QS, B, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, B-Syl, B-2, Em-Some, Em-Right)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, B-2, B-Syl, Em-Some, Em-Right)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, B-Syl, Em-Some, Em-Right, B-2)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, B-2)
(QS, Ft Hd Left, B, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(QS, Ft Hd Left, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(QS, Ft Hd Left, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(B, QS, Ft Hd Left, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-Syl, B-2, Em-Some, Em-Right)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, B-Syl, Em-Some, Em-Right)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, B-Syl, Em-Some, Em-Right, B-2)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, B-2)
(B, QS, Ft Hd Left, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)

(B, Ft Hd Left, QS, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B-Syl, B-2, Em-Some, Em-Right)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B-2, B-Syl, Em-Some, Em-Right)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B-2, Em-Some, Em-Right, B-Syl)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, B-Syl, Em-Some, Em-Right, B-2)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-Syl, B-2)
(B, Ft Hd Left, QS, Ft Dir Rt, QS-VC-Heavy, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(Ft Hd Left, QS, QS-VC-Heavy, B, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, B-Syl, B-2, Em-Some, Em-Right)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, B-Syl, Em-Some, Em-Right)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, B-2, Em-Some, Em-Right, B-Syl)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, B-Syl, Em-Some, Em-Right, B-2)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-Syl, B-2)
(Ft Hd Left, QS, B, QS-VC-Heavy, Ft Dir Rt, Em-Some, Em-Right, B-2, B-Syl)

(A3) Viable Parameter-Setting Orders for the Cues Method, Relativize-Against-Potential

No viable orders.

(A4) Viable Parameter-Setting Orders for the Parsing Method, Relativize-Against-Potential

Same set as (A2): parsing method and relativize-against-all approach.

Bibliography

Akhtar, N., Callanan, M., Pullum, G., & Scholz, B. (2004). Learning antecedents for anaphoric *one*. *Cognition, 93,* 141-145.

Archibald, J. (1992). Adult abilities in L2 speech: evidence from stress. In J. Leather & A. James, eds. *New Sounds 92: Proceedings of the 1992 Amsterdam Symposium on the Acquisition of Second Language Speech*: 1-16. Amsterdam: University of Amsterdam Press.

Archibald, J. (1998). *Second Language Phonology*. Amsterdam: Benjamins.

Bailey, C-J. (1973). *Variation and Linguistic Theory.* Washington, DC: Center for Applied Linguistics.

Baker, C. L. (1979). *Syntactic theory and the projection problem*. Linguistic Inquiry, *10*, 533-81.

Baker, M. (2001). *The Atoms of Language: The Mind's Hidden Rules of Grammar.* New York, NY: Basic Books.

Baker, M. (2005). Mapping the Terrain of Language Learning. *Language Learning and Development, 1,* 93-129.

Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Berwick, R. and Weinberg, A. (1984). *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. Cambridge, MA: MIT Press.

Bock, J. & Kroch, A. (1989). The Isolability of Syntactic Processing. In G. Carlson, & M. Tannenhaus (Eds.), *Linguistic Structure in Language Processing*. Boston: Kluwer.

Bonatti, L.L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on Statistical Computations. *Psychological Science, 16*(6), 451-9.

Booth, A. & Waxman, S. (2003). Mapping words to the world in infancy: on the evolution of expectations for nouns and adjectives. *Journal of Cognition and Development*: *4*(3), 357-381.

Briscoe, T. (1999). The Acquisition of Grammar in an Evolving Population of Language Agents, *Electronic Transactions on Artificial Intelligence, 3*.

Briscoe, T. (2000). An evolutionary approach to (logistic-like) language change. Ms., University of Cambridge.

Canavan, A., Graff, D., and Zipperlen, G. (1997). *CALLHOME American English Speech*. Linguistic Data Consortium: Philadelphia, PA.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*, New York: Harper and Row.

Cinque, G. (1999). *Adverbs and Functional Heads: A Cross-linguistic Perspective*. Oxford: Oxford University Press.

Clahsen, H. (1986). Verbal inflections in German child language: acquisition of agreement markings and the functions they encode. *Linguistics, 24,* 79-121.

Clark, R. (1992). The Selection of Syntactic Knowledge. *Language Acquisition*, *2*(2), 83-149.

Clark, R. (1994). Kolmogorov complexity and the information content of parameters. *IRCS Report 94-17*. Institute for Research in Cognitive Science, University of

Pennsylvania.

Clark, R., & Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry, 24,* 299-345.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement and uncertainty, *Cognition, 58,* 1-73.

Dale, P.S. & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127.

Dresher, E. (1994). Acquiring stress systems. In Ristad, E. (ed.), *Language computations*, Providence, RI: AMS, 71-92.

Dresher, E. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, *30*, 27-67.

Dresher, E. & Lahiri, A. (2003). Main Stress Left in Early Middle English, in Fortescue, M, Skafte Jensen, E., Mogensen, J. and Schøsler, L. (eds.), *Historical Linguistics 2003. Selected Papers from the 16th International Conference on Historical Linguistics. Copenhagen. 10-15 August 2003*. Amsterdam: John Benjamins.

Fodor, J. D. (1998a). Unambiguous Triggers. *Linguistic Inquiry, 29*, 1-36.

Fodor, J. D. (1998b). Parsing to Learn. *Journal of Psycholinguistic Research, 27*(3), 339-374.

Fodor, J. (1998c). Learning to parse? *Journal of Psycholinguistic Research*, *27*, 285-319.

Foraker, S., Regier, T., Khetarpal, A., Perfors, A., and Tenenbaum, J. (in press). Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. To appear in *Proceedings of the 2007 Cognitive Science conference.*

Gallistel, C.R. (2001). Mental Representations, Psychology of. In *Encyclopedia of the social and behavioral sciences*. New York: Elsevier.

Gerken, L. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition, 93,* B89-B96.

Gerken, L. (2006). Decision, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67-B74.

Giles, H. and Powesland, P. (1975). *Speech Styles and Social Evaluation.* London: Academic Press.

Goldsmith, J. & O'Brien, J. (2006). Learning Inflectional Classes. *Language Learning and Development*, *2*(4), 219-250.

Golinkoff, R.M., Hirsh-Pasek, K., Cauley, K.M., Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language, 14,* 23-45.

Halle, M. & Idsardi, W. (1995). General Properties of Stress and Metrical Structure, in Goldsmith, J. (ed.), *The Handbook of Phonological Theory*, Cambridge, MA & Oxford: Blackwell Publishers, 403-443.

Halle, M. & Vergnaud, J-R. (1978). *Metrical structures in phonology*. Ms., Cambridge: MA.

Halle, M. and Vergnaud, J-R. (1987). *An Essay on Stress*. Cambridge, MA: MIT Press.

Hamburger, H. & Crain, S. (1984). Acquisition of cognitive compiling. *Cognition*, *17*, 85-136.

Harris, J. (1983). *Syllable Structure and Stress in Spanish: A Nonlinear Analysis.*

Cambridge: MIT Press.

Hayes, B. (1980). *A Metrical Theory of Stress Rules.* Ph.D. Dissertation, M.I.T.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

Hornstein, N., & Lightfoot, D. (1981). *Explanation in linguistics: the logical problem of language acquisition*. London: Longmans.

Houston, D., Jusczyk, P, Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review, 7*(3), 504-509.

Houston, D., Santelmann, L., & Jusczyk, P. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language & Cognitive Processes, 19*(1), 97-136.

Hudson Kam, C.L., & Newport, E.L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*, 151-195.

Idsardi, W. (1992). The Computation of Prosody. Doctoral dissertation, MIT, Cambridge, MA.

Jusczyk, P. & Aslin, R. (1995). Infants' detection of the sound pattern of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Jusczyk, P. Cutler, A., & Redanz, N. (1993). Infants' sensitivity to predominant word stress patterns in English. *Child Development, 64*, 675-687.

Jusczyk, P., Houston, D., Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*(3-4), 159-207.

Jurafsky, D. & Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kager, R. (1995). "The metrical theory of word stress", in Goldsmith, J. (ed.), *The Handbook of Phonological Theory*, 367-402. Oxford: Basil Blackwell.

Kibler, W. (1984). *An Introduction to Old French.* New York: Modern Language Association of America.

Koenigsberger, H.G., & Briggs, A. (1987). *Medieval Europe, 400-1500.* Longman: New York.

Kroch, A., & Taylor, A. (1997). Verb Movement in Old and Middle English: Dialect Variation & Language Contact. In van Kemenade, A. & Vincent, N. (eds.), *Parameters of Morphosyntactic Change.* Cambridge: Cambridge University Press, 297-325.

Kroch, A., & Taylor, A. (2000). The Penn-Helsinki parsed corpus of Middle English. Philadelphia: Department of Linguistics, University of Pennsylvania, 2nd edn. Accessible via http://www.ling.upenn.edu/mideng.

Lasnik, H. (1987). A note on indirect negative evidence. *UConn Working Papers in Linguistics, 1,* 19-26.

Legate, J. & Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review, 19,* 151-162.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition, 89,* B65-B73.

Lidz, J. & Waxman, S. (2004). Reaffirming the poverty of the stimulus argument: a reply to the replies. *Cognition, 93,* 157-165.

Lightfoot, D. (1982). *The Language Lottery: Toward a Biology of Grammars,* Cambridge, MA: MIT Press.

Lightfoot, D. (1991). *How to Set Parameters: arguments from language change*, Cambridge, MA: MIT Press.

Lightfoot, D. (1999). *The Development of Language: Acquisition, Change, and Evolution.* Oxford: Blackwell.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The MIT Press.

Manzini, R. and Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry* 18.3, 413-444.

Mehler, J., Juszcyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*(2), 143-178.

Morgan. (1986). *From Simple Input to Complex Grammar.* Cambridge, MA: MIT Press.

Nazzi, T., Jusczyk, P., & Johnson, E. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory & Language, 43*(1), 1-19.

Newport, E. & Aslin, R. (2004). Learning at a distance: Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127-162.

Neyman, J. & Pearson, E. (1928) On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, *20A(1/2)*: 175-240.

Niyogi, P., & Berwick, R. (1995). *The logical problem of language change.* AI-Memo 1516, Artificial Intelligence Laboratory, MIT.

Niyogi, P., & Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition, 61,* 161-193.

Niyogi, P., & Berwick, R. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy, 20,* 697-719.

Osgood, C., & Sebeok, T. (1954). Psycholinguistics: a survey of theory and research problems. *Journal of Abnormal and Social Psychology*, *49,* 1-203.

Pearl, J. (1996). Decision making under uncertainty. *ACM Computing Surveys (CSUR)*, 28.1, 89-92.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the Stimulus? A rational approach. *28th Annual Conference of the Cognitive Science Socity.* Vancouver, British Columbia.

Pierce. A. (1992). *Language Acquisition and Syntactic Theory: A Comparative Analysis of French and English Child Grammars*. Boston, MA: Kluwer Academic.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7*, 217-283.

Pintzuk, S. (2002). Verb-Object Order in Old English: Variation as Grammatical Competition. *Syntactic Effects of Morphological Change.* Oxford: Oxford University Press.

Regier, T. & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition, 93,* 147-155.

Rizzi, L. (1997). The Fine Structure of the Left Periphery. In L. Haegeman (ed.), *Elements of Grammar: Handbook of Generative Syntax*. Dordrecht: Kluwer, 281-337.

Rizzi, L. (2004). Locality and the Left Periphery. In A. Belletti (ed.), *Structures and Beyond. The Cartography of Syntactic Structures*, *Volume 3*. Oxford: Oxford University Press, 223-251.

Saffran, J., Aslin, R., and Newport, L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J., Newport, L., and Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606-621.

Sakas, W.G. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* 415-422.

Sakas, W.G. & Fodor, J.D. (2001). The structural triggers learner. In S. Bertolo (ed.) *Language Acquisition and Learnability*, Cambridge University Press, Cambridge, UK.

Sakas, W. & Nishimoto, E. (2002). Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Ms., CUNY: New York.

Shannon, C. (1948). A mathematical theory of communication,' *Bell System Technical Journal, 27*, 379-423 and 623-656.

Skinner, B. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Spelke, E. S. (1979). Perceiving Bimodally Specified Events in Infancy. *Developmental Psychology, 15 (6)*, pp. 626-636.

Staddon, J.E.R. (1988).Learning as Inference. In Bolles, R. and Beecher, M. (eds.), *Evolution and Learning*, Hillside, NJ: Lawrence Erlbaum.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). The York-Toronto-Helsinki parsed corpus of Old English. York, UK: Department of Language and Linguistic Science, University of York. Available through the Oxford Text Archive.

Tenenbaum, J. & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-640.

Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309-318.

Thompson, S. & Newport, L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3,* 1-42.

Thornton, R. & Crain, S. (1994). Successful cyclic movement. In Hoekstra, T. & Schwartz, B. (eds.), *Language Acquisition Studies in Generative Grammar*, John Benjamins, 215-253.

Thornton, R. & Crain, S. (1999). Levels of representation in child grammar. *The Linguistic Review, 16,* 81-123.

Tirumalesh, K.V. (1996) "Topic and Focus in Kannada: Implications for Word Order," South Asian Language Review, 6.1, 25-48.

Valian, V. (1990). Null subjects: A problem for parameter setting models of language acquisition. *Cognition*, *35*, 105-122.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40,* 21-82.

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W. & Malkiel, Y. (eds.), *Directions for Historical Linguistics.* Austin: University of Texas Press.

Wexler, K. & Culicover, P. (1980). *Formal Principles of Language Acquisition.* Cambridge, MA: MIT Press.

Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, *20(1)*, 6-11.

Yang, C. (2000). Internal and external forces in language change. *Language Variation and Change, 12*, 231-250.

Yang, C. (2002). *Knowledge and Learning in Natural Language.* Oxford: Oxford University Press.

Yang, C. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Science, 8(10),* 451-456.