

Running Head:

When Probabilistic Learning Is Not Enough

Title:

When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology

Author:

Lisa S. Pearl

Department of Cognitive Sciences

University of California, Irvine

lpearl@uci.edu

Abstract

Parametric systems have been proposed as models of how humans represent knowledge about language, motivated in part as a way to explain children's rapid acquisition of linguistic knowledge. Given this, it seems reasonable to examine if children with knowledge of parameters could in fact acquire the adult system from the data available to them. That is, we explore an argument from acquisition for this knowledge representation. We use the English metrical phonology system as a non-trivial case study, and test several computational models of unbiased probabilistic learners. Special attention is given to the modeled learners' input and the psychological plausibility of the model components in order to consider the learning problem from the perspective of children acquiring their native language. We find that such cognitively-inspired unbiased probabilistic learners uniformly fail to acquire the English grammar proposed in recent metrical studies from English child-directed speech, suggesting that probabilistic learning alone is insufficient to acquire the correct grammar when using this parametric knowledge representation. Several potential sources of this failure are discussed, along with their implications for the parametric knowledge representation and the trajectory of acquisition for English metrical phonology.

1. Introduction: Knowledge representations

Acquisition is not so easy if we believe children are acquiring a generative system for their language (e.g., Chomsky 1981, Halle & Vergnaud 1987, Hayes 1995, Tesar & Smolensky 2000, Prince & Smolensky 2004, Heinz 2007, among many others), rather than less abstract representations of the data they encounter (e.g., Daelemans, Gillis, & Durieux 1994, Goldberg 1995, Gillis, Daelemans, & Durieux 2000, Tomasello 2006, among many others). The idea of a complex linguistic system that varies over a limited number of dimensions (often called *parameters* or *constraints*) serves a dual purpose. First, it is used to explain the constrained variation seen in adult languages cross-linguistically within some specific domain (e.g., metrical phonology (Halle & Vergnaud 1987, Hayes 1995) or syntax (Chomsky 1981)). That is, it is often argued that it is surprising to see such limited variation if there is no common underlying system that humans are using. Motivation for many theoretical representations comes from examining adult language knowledge and attempting to describe a system that can compactly capture the limited variation observed. For example, Hayes (1995, p.55) states that the requirements of a successful theory of stress are that it is “well defined”, “maximally restrictive”, and is “capable of describing all the stress systems of the world’s languages”. Another motivation is tied explicitly to acquisition. Specifically, systematic linguistic knowledge is used to explain how children come to know what they do about language so quickly (e.g., Chomsky 1981, Dresher 1999). In this case, there is often an implicit assumption that if children have prior knowledge of the linguistic system, acquisition from the available data will be easier (e.g., see Crain & Pietroski (2002)’s discussion on “why language acquisition is a snap” for aspects of syntax and semantics).

However, it is often unclear how exactly this acquisition process works, given the complexity of the systems proposed and the data children learn from. In this paper, we explore the second motivation described above – the argument from acquisition – for a parametric knowledge representation, which is a knowledge representation that has been motivated mainly, to my knowledge, by cross-linguistic comparisons of adult knowledge.

1.1. Acquiring complex systems

The proposal that children build complex systems from the available data is perhaps not too unreasonable – there is evidence that children search for linguistic generalizations in the available data, even when generalization is not required in order for children to effectively use the language. An example of this situation is metrical phonology, where the number of potential stress contours for any given word form is finite. Children could simply memorize the appropriate stress contour for a word, but there are studies suggesting that children nonetheless seem to look for generalizations. For example, Hochberg (1988) shows that Spanish children learning the metrical phonology of their language learn rules for assigning stress even when they could simply memorize the stress patterns on a word-by-word basis. Nouveau (1994) demonstrates similar behavior in Dutch children. Also, work by Kehoe (1997) demonstrates that English children produce greater numbers of errors on exceptional word forms, supporting a systematic representation of stress.

To build the correct system rapidly, it is hypothesized that children have prior knowledge of the parameters of variation available in the complex system (e.g., Chomsky 1981, Dresher 1999). Without this prior knowledge, it would be difficult to decide the

relevant points of variation (henceforth *parameters*) among all the potential ways the system might vary, and also to decide the correct values for those parameters. This is why having knowledge of the parametric system beforehand makes acquisition easier.

Yet how can we test this proposal? If we have a knowledge representation specified and we have some idea what the data are like that children use, we can create computational models that are grounded both theoretically and empirically, and thus we can concretely examine questions about the acquisition process associated with a particular knowledge representation. Specifically, within a model, we can precisely define the acquisition process and see the results of that definition for acquisition.

Notably, acquisition is more constrained than the general learnability problems often considered in computational learning. For instance, acquisition includes limitations on the type of input the learner receives, the duration of learning, and the processing capabilities of the learner. Here, we will model the acquisition of a parametric system of metrical phonology, using child-directed speech data as input, and keeping in mind restrictions on the time course of acquisition and children's cognitive limitations. If we believe the model reasonably reflects the process of acquisition in children, the results that come from the manipulations of this process in the model inform us about the nature of this process in children. From an empirical standpoint, some manipulations we can easily do within a modeled learner are more difficult to do with children – such as controlling the hypotheses they entertain (in this case, about different parameters), the data they learn from, and the way they change beliefs in competing hypotheses. For this reason, the results from modeling can be particularly informative about these aspects of

acquisition, with respect to both what will and what will not work for a particular knowledge representation.

1.2. Parametric metrical phonology

A parametric system of metrical phonology is one proposal to explain the constrained variation seen in the world's stress systems (Halle & Vergnaud 1987, Hayes 1995). To acquire a parametric system, children must view the encountered data as the output of that generative system and deconstruct those data in order to identify the parameter values involved. If we consider metrical phonology, the output is the stress contour associated with a given word, including the basic division into stressed and unstressed syllables. Suppose a child encounters the word unicorn (stressed syllables will be indicated by underlining henceforth), which has the stress contour [stressed unstressed stressed]. Even if the child is primed to acquire a parametric system, the task is very difficult without knowing the relevant parameters. A parameter could be any variable present in the child's linguistic or non-linguistic experience; for instance, the child might consider (a) if the individual segments of the word matter (e.g., *u*, *n*, *r*), (b) if the individual syllables matter (e.g., *u*, *corn*), (c) if rhyming matters (e.g., *u* does not rhyme with *corn*), (d) if the speaker's rate of speech matters (e.g., fast vs. normal speech), (e) if the speaker's gender matters, (e.g., female vs. male speech), and so on. Knowing which parameters are relevant significantly constrains the child's hypothesis space of language systems (henceforth *grammars*). In addition, knowing the range of values these parameters can have also reduces the hypothesis space.

Still, even with this prior knowledge, the hypothesis space of possible grammars can be quite large as it grows exponentially with the number of parameters. For example, suppose the child is aware of n binary parameters. Then, there are 2^n possible grammars in the hypothesis space. Even if n is small (say 20), this can lead to a very large number of potential grammars ($2^{20} = 1,048,576$). Children still must choose among a fairly large number of hypotheses. So, having a parametric system defined a priori certainly does not solve the acquisition problem by itself.

In addition, the hypothesized cross-linguistic parameters often interact, so the observable data are ambiguous between a number of available grammars (Clark 1994, among others). Consider, for example, a stress contour such as [stressed unstressed stressed] in a word like *afternoon*. In (1), we see just a few of the analyses generated from grammars that can yield this stress contour. Syllables are either undifferentiated (S), or divided into Light (L) and Heavy (H) syllables, according to the syllable's structure. Larger units called metrical feet (indicated by parentheses (...)) are then formed that are made up of one or more syllables, and stress is assigned inside each metrical foot.

(1) Generative grammar analyses compatible with the stress contour of *afternoon*

- | | | |
|--------------------|--------------------|--------------------|
| (a) (S S) (S) | (b) (L L) (H) | (c) (L) (L H) |
| <u>af</u> ter noon | <u>af</u> ter noon | <u>af</u> ter noon |

Metrical phonology system parameters include which syllables are contained in metrical feet, how large metrical feet are, and which syllables are stressed inside metrical

feet. Even if these parameters are known already, it can be difficult to determine which parameter values combined to yield the observed stress contour. So, even with this prior knowledge, the acquisition problem is again not really solved, even if there are relatively few parameters involved. The acquirability of the correct grammar from the available data is still an open question.

1.3. The present case study: English metrical phonology

Here, we examine the acquirability of the English grammar defined with respect to a specific parametric system of metrical phonology (Dresher (1999), Dresher & Kaye (1990), Halle & Vergnaud (1987)), based on the input available to English children. Note that for the remainder of this paper, whenever the terms “English grammar” or “adult grammar” are used, they refer to the adult English grammar derived from these metrical phonology studies and do not refer to how the English grammar may be defined for other possible knowledge representations. As such, the results reported here impact this specific knowledge representation only.

Prior computational modeling work (Pearl 2008, 2009) has suggested that this adult grammar is acquirable from English child-directed speech as long as the child (1) employs a selective learning bias, (2) is sensitive to the probability distributions in the data, and (3) sets parameters in particular orders (see section 5 and Appendix A3 for more discussion of this learning strategy). Since this system is acquirable from realistic data (even if additional knowledge is required in the form of the learning bias and the parameter-setting orders), we have a mark in favor of this proposal of knowledge representation. However, an open question is whether that additional knowledge is

necessary for a child who can leverage the information available in probability distributions. That is, can children who simply track probabilistic information also acquire the correct grammar?

The metrical phonology system is a tractable case study to explore as the hypothesis space can be explicitly defined by a reasonably small number of parameters drawn from cross-linguistic research. However, these parameters interact and thus make identifying the parameter value responsible for a given stress contour non-trivial. The instantiation of the metrical phonology parameters used in this case study are adapted from the parameterization used in Dresher (1999), which is based on the system in Dresher & Kaye (1990) that draws from Halle & Vergnaud (1987). Also, a metrical phonology system very similar to the one here has been used to study the acquisition of stress by second-language speakers (Archibald 1992). In this system, there are five main binary parameters and four binary sub-parameters, yielding 156 legal grammars in the hypothesis space (this is due to some dependencies between the parameters, discussed in more detail in section 2 below). The resultant grammars concern only whether syllables are stressed or unstressed, and not how much stress syllables receive compared to other syllables.¹ Moreover, these grammars do not describe interactions with the morphology

¹ Once this first basic distinction is made, children can then decide where the main stress of the word lies, since we will assume that main stress assignment depends on first knowing which syllables are stressed. In grid theory, as described in Dresher (1999) which draws from Prince (1983), stressed vs. unstressed corresponds to distinctions made on Line 1, while main stress corresponds to distinctions made on Line 2, i.e., main stress is an additional level of representation beyond the distinction between stressed and unstressed. However, see Prince (1983), Hayes (1995), and Prince & Smolensky (2004) for discussion of this assumption, which is controversial for different levels of metrical representation within languages.

system, due to considerations of the child's likely initial knowledge state when acquiring the metrical phonology system. Specifically, children may not have hypothesized the connection between the morphology system and the metrical phonology system. Kehoe (1998) suggests that children may already know several parameter values of the English system by 22 months. It seems unlikely that children at this age have extensive knowledge of their language's morphology, since Brown (1973) suggests that it is just after 36 months that they begin to use morphological endings with some regularity. We thus proceed with the following premise: the child's first hypothesis about the metrical phonology system is that it is autonomous, and does not interact with other systems. Given this, the child first attempts to identify the grammar in the hypothesis space that is most compatible with the available data, perhaps noting that there are exceptions to this grammar once the grammar has been identified. Later, the child may recognize that some exceptions are systematic, and can be captured by considering interactions with the morphology system.

It is important to note that the metrical phonology system considered here, while not a full adult system, is still significantly more complex than parametric systems explored in some prior computational modeling studies, which involved at most three interacting parameters (Gibson & Wexler 1994, Niyogi & Berwick 1996, Pearl & Weinberg 2007). Previous studies that have examined parametric systems of similar complexity to the one considered here have often not used child-directed speech as input when assessing the system's acquirability (Dresher 1999, Dresher & Kaye 1990, Fodor & Sakas 2004, Sakas 2003, Sakas & Nishimoto 2002, among others). To address this, the modeled learners here use a data set drawn from CHILDES (MacWhinney 2000) as input that contains

both the forms children are likely to encounter and the frequencies at which they will encounter these forms.

In the remainder of the paper, we review the parameters of the metrical phonology system under consideration, and then describe the analysis of the English child-directed speech data. Following that, we present several cognitively-inspired unbiased probabilistic learners that attempt to acquire the adult English grammar from data distributions estimated from English child-directed speech. Surprisingly, we will find that they fail to acquire the correct grammar with any reliability. Following that, we identify the source of their failure and find it is due to the definition of the acquisition task – specifically, the definition of the target state and the data distributions in the English child-directed speech data. For this reason, *no* unbiased probabilistic model would succeed at this task, even if it was not constrained in the way our cognitively-inspired modeled learners are. We conclude with discussion of implications for this parametric knowledge representation and the acquisition process.

2. The parameters of the system

A sample metrical phonology analysis using the English grammar is shown for *octopus* in (2). The word is divided into syllables (*oc, to, pus*), which are then classified according to syllable structure as either (L)ight or (H)eavy. The rightmost syllable (*pus*) is extrametrical (indicated by angle brackets <...>), and so not contained in a metrical foot. The metrical foot spans two syllables (*oc, to*), and the leftmost syllable within the foot (*oc*) is stressed. This leads to the observable stress contour: *octopus*.

(2) metrical phonology analysis for *octopus*

(H L) <H>
oc to pus

As we can see, many parameters combine to produce the word's stress contour in this system. We will now briefly step through the various parameters involved (drawn from Dresher (1999) and Dresher & Kaye (1990)). For a more detailed description of each of the parameters and their interactions with each other, see Dresher & Kaye (1990), Dresher (1999), and Pearl (2007).

One parameter, quantity sensitivity, concerns whether all syllables are identical, or differentiated by syllable rime weight for the purposes of stress assignment. The rime consists of the nucleus and coda only, so this definition of weight is insensitive to the syllable onset (e.g., *en = ten = sten = stren*). A language could be *quantity sensitive* (QS), so that syllables are differentiated into (H)heavy and (L)ight syllables. Long vowel syllables with or without codas (VV(C)) are Heavy, short vowel syllables (V) are Light, and short vowel syllables with codas (VC) are either Light (QS-VC-L) or Heavy (QS-VC-H). In contrast, if the language is *quantity insensitive* (QI), all syllables are identical (represented below as 'S'). Both kinds of analyses are shown in (3) for *unicorn*.

(3) QS and QI analyses of *unicorn*

QS analysis	H	L	L/H	QI analysis	S	S	S
syllable rime	VV	V	VC				
syllable structure	VV	CV	CVCC				

syllables

u ni corn

u ni corn

Syllables classified as Heavy should be more prominent than syllables classified as Light, and one way this prominence can be expressed is by receiving stress (following the assumption in Dresher (1999) that draws from Prince (1983)). However, sometimes Heavy syllables are not stressed. In the current parametric system, this can be due to another parameter, extrametricality, which concerns whether all syllables of the word are contained in metrical feet. Only syllables contained in metrical feet receive stress, so an excluded Heavy syllable will not be stressed. In languages with extrametricality (Em-Some), either the leftmost syllable (Em-Left) or the rightmost syllable (Em-Right) is excluded. In contrast, languages without extrametricality (Em-None) have all syllables included in metrical feet. Example (4a) shows Em-Some analyses for *giraffe* and *octopus*, while (4b) shows an Em-None analysis for *afternoon*. Note that these are not necessarily the analyses believed to hold for English – they are simply analyses compatible with the observable stress contour.

(4a) Em-Some analyses, QS, QS-VC-H

	Em-Left		Em-Right
syllable class	<L>	(H)	(H L) <H>
syllable rime	V	VC	VC V VC
syllables	<i>gi</i>	<i>raffe</i>	<i>oc to pus</i>

(4b) An Em-None analysis, QS, QS-VC-L

syllable class	(L L)	(H)
----------------	-------	-----

syllable rime	VC VC VVC
syllables	<u>af</u> ter <u>noon</u>

Once the syllables to be included in metrical feet are known, metrical feet can be constructed. The foot directionality parameter controls which side of the word metrical foot construction begins at, the left (Ft-Dir-Left) or the right (Ft-Dir-Rt). Examples of both options are shown in (5).

(5a) Start metrical foot construction from the left (Ft-Dir-Left): (L L H

(5b) Start metrical foot construction from the right (Ft-Dir-Rt): L L H)

Then, the size of metrical feet must be determined by the boundedness parameter. An unbounded (Unb) language has no arbitrary limit on foot size; a metrical foot is only closed upon encountering a Heavy syllable or the edge of the word. If there are no Heavy syllables or the syllables are undifferentiated (because the QI value is used), then the metrical foot encompasses all the non-extrametrical syllables in the word. Some example Unb analyses are shown in (6).²

(6) Unb analyses

² Note that the assignment of stress has not yet taken place (this does not occur until the foot headedness parameter discussed below). However, as Heavy syllables should be stressed, the unbounded analyses shown in (6) end up with foot directionality being the same direction as foot headedness, in order to create a stress assignment where word-internal heavy syllables are stressed. This will not necessarily be true for the Bounded analyses, as shown below in (7).

(a) Building feet from the left (Ft-Dir-Left), QS, Em-None

(Step 1) (L L L H L) begin

(Step 2) (L L L) (H L) heavy syllable encountered

(Step 3) (L L L) (H L) end

(b) Building feet from the right (Ft-Dir-Rt), QS, Em-None

(Step 1) (L L L H L) begin

(Step 2) (L L L H) (L) heavy syllable encountered

(Step 3) (L L L H) (L) end

(c) Building feet from the left (Ft-Dir-Left) with all light syllables, QS, Em-None

(Step 1) (L L L L L) begin

(Step 2) (L L L L L) end

(d) Building feet from the right (Ft-Dir-Rt) with undifferentiated syllables (QI) and Em-None

(Step 1) (S S S S S) begin

(Step 2) (S S S S S) end

The alternative is for metrical feet to be Bounded (B), and so to be no larger than a specific size. A metrical foot can be either two units (B-2) or three units (B-3); units are either syllables (B-Syl) or sub-syllabic units called moras (B-Mor) that are determined by the syllable's weight (Heavy syllables are two moras while Light syllables are one). Only if the word edge is reached can metrical feet deviate from this size (by being smaller than this size). Example (7) demonstrates different bounded analyses, with various combinations of these parameter values.

B-2/3, and B-Syl/Mor) yield 156 legal grammars in the hypothesis space.³ Since these parameters interact, a change to any one of their values could non-trivially change the stress contour. For example, consider (9), where changing the extrametricality parameter from Em-Right to Em-Left causes the entire stress contour to become its inverse (i.e., all syllables that were previously stressed are now unstressed, and all syllables that were previously unstressed are now stressed).

(9) Consequences of changing a single parameter for a four syllable sequence

(a) QI, Em-Some, **Em-Right**, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left

(S S) (S) <S> → S S S S

(b) QI, Em-Some, **Em-Left**, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left

<S> (S S) (S) → S S S S

Due to parameter interaction, it may be difficult for a child to determine if a particular parameter value is responsible for generating the correct stress contour. This has been called the Credit Problem (Dresher 1999), and is the result of data ambiguity. For example, consider two grammars that the word *cucumber* is compatible with (10). These two grammars share no parameter values whatsoever in common, making it difficult to determine which parameter values should be credited with correctly generating the observed stress contour.

³ Note that this is less than the 180 possible grammars, as grammars including both quantity insensitivity (QI) and bounded moraic (B-Mor) are ruled out - counting by moras requires treating syllables as Heavy and Light (i.e., using the QS value).

(10) Two grammars *cucumber* is compatible with

(a) QI, Em-None, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Left

syllable class (S) (S) (S)

syllables cu cum *ber*

(b) QS, QS-VC-H, Em-Some, Em-Right, Ft-Dir-Left, Unb, Ft-Hd-Rt

syllable class (H) (H) <H>

syllables cu cum *ber*

3. English

The particular language considered in this modeling study is English, which has the following parameter values (following Dresher (1999), who draws from the analysis in Halle & Vergnaud (1987)): QS, QS-VC-H, Em-Some, Em-Right, Ft-Dir-Rt, B, B-2, B-Syl, and Ft-Hd-Left. There are several reasons English was chosen as the target language. First, English child-directed speech data are very ambiguous with respect to the 156 grammars in the hypothesis space, making the acquisition problem non-trivial. Second, there are numerous irregular data – we can easily see this if we examine syllable types, where a syllable type is defined as a sequence of syllables specified by their rime (e.g., VC (closed)-VC (closed) is a 2-syllable syllable type, and corresponds to many different 2-syllable vocabulary items). A survey of the Brent corpus of English speech directed to children between the ages of eight and fifteen months (Brent & Siskind 2001) reveals 174 separate syllable types for words of two or more syllables. Of these 174, 85

have more than one stress contour associated with them. For example, the VC-VC syllable type includes the words *herself*, *answer*, and *somewhere*, which all have different stress contours. Since a grammar generates a stress contour based on the syllable type, this means that no one grammar can be compatible with these 85 syllable types. Instead, a grammar may be compatible with one of the associated stress contours for a given syllable type of this set of 85 but will, by definition, not be compatible with any others that are associated with it. In the example above, no one grammar would be able to produce the contours associated with all three words – at best, a grammar would produce a contour compatible with one word (e.g., words like *answer*) and incompatible with the other two (e.g., words like *herself* and *somewhere*). Thus, the English data set is noisy in this respect, and certainly complicates acquisition of a generative system. We should note that the task for the learners here is to select the adult English grammar, given the English child-directed speech data, but not necessarily to note which words are exceptions to this grammar. The noting of exceptions would presumably occur after the child has figured out which grammar to choose.⁴ The child would then identify exceptional data points as ones that the selected grammar cannot analyze.

⁴ However, one can imagine that a child might be able to begin noting exceptions for some of the parameter values. For example, if the child realizes the grammar uses the quantity sensitive (QS) value, the child might be able to recognize words containing internal unstressed heavy syllables (e.g., the *y* in *ponytail*) as being incompatible with this parameter value. These words could then be ignored. It should be noted that this requires some additional processing/reasoning on the child's part, such as realizing that this observable structural pattern is incompatible with the quantity sensitive value. This is similar to the knowledge required to recognize unambiguous data, in this case for the quantity insensitive (QI) value. Depending on the parameter value, this may be easy or rather difficult. See Appendix A3 for discussion.

Still, we should not give up hope completely on a generative system. While there obviously must be some way to deal with these exceptional data, a grammar that can reliably cover a large portion of the data is still a useful grammar for children to have. Also, this situation is not too unusual for metrical acquisition data; for example, Daelemans *et al.* (1994) note that 20% of the Dutch data they consider are irregular according to a generally accepted metrical analysis and so must be dealt with in terms of idiosyncratic marking.

A third reason for using English as our case study is that previous computational modeling research (Pearl 2008, 2009) has found that the adult English grammar can be acquired in this parametric system from child-directed English speech data if the child has a bias to learn only from unambiguous data (Fodor 1998a, Dresher 1999, Lightfoot 1999, Pearl & Weinberg 2007) and the parameters are acquired in a particular order. Given a possible way to succeed using a bias, we can now explore whether successful acquisition for this difficult case specifically requires a bias or is merely aided by it. If unbiased learners are successful, we know that a bias – while helpful – is not strictly necessary. This is attractive as the successful bias found previously required prior knowledge or potentially intensive processing to implement (see Pearl (2007, 2008) for details), in addition to restrictions on the order in which parameter values could be set. However, if unbiased learners are unsuccessful, we can examine why they fail and whether the problem that afflicts these modeled learners is model-specific or endemic to all unbiased models. Fourth, and finally, numerous English child-directed speech samples are available through CHILDES (MacWhinney 2000), so realistic estimates of

the data distributions children encounter can be made. Thus, our argument from acquisition for this parametric knowledge representation can be empirically grounded.

4. The model

4.1. The learner's input

The learner's input was derived from the distributions of words and their associated stress contours in English child-directed speech samples. The Bernstein-Ratner corpus (Bernstein Ratner 1984) and the Brent corpus (Brent & Siskind 2001) were selected from the CHILDES database (MacWhinney 2000) because they contain speech to children between the ages of eight months and two years old. This age range was estimated as the time period when children might be beginning to set the parameters of the metrical phonology system under consideration, given that several parameters of this system may be known by 28 months (Kehoe 1998). The Bernstein-Ratner corpus consists of recordings of nine child-mother dyads during play sessions, with the children ranging in age between 1;1 and 1;11. The Brent corpus consists of sixteen sets of mothers speaking to preverbal infants between the ages of 0;8 and 1;3. In total, this yielded 540,505 words of orthographically transcribed child-directed speech, consisting of 8,093 word types. For the most part, words were defined as strings of text surrounded by space, though there were some exceptions such as words connected by +, like *nightie+night*. A child's syllabification of these words was estimated by using the MRC Psycholinguistic Database (Wilson 1988), and the associated stress contour was estimated by referencing the CALLHOME American English Lexicon (Canavan *et al.* 1997). Words not present in these two databases were given a syllabification/pronunciation consistent with the

conventions in the two databases – such words were usually child-register words, e.g. *booboo*. See Appendix A1 for a detailed summary of the corpus.

The simulated learners learned from 1,666,667 words sampled from this data set, as this was the estimated number of tokens children would hear in a six month period⁵, based on the estimates for word tokens heard in a three year period (10 million) in Akhtar et al. (2004) (citing Hart & Risley (1995)).

4.2. The modeling framework

All the models described below fit into a very general modeling framework involving three components: a definition of the hypothesis space, a definition of the data intake, and a definition of the update procedure (Pearl 2007, Pearl & Lidz 2009). The hypothesis space here is defined in terms of competing grammars, similar to other previous modeling work (Clark 1992, Gibson & Wexler 1994, Niyogi & Berwick 1996, Sakas & Fodor 2001, Sakas & Nishimoto 2002, Yang 2002, Sakas 2003, Fodor & Sakas 2004, Pearl & Weinberg 2007, Pearl 2008). The data intake is all the available input, which is derived from the frequencies in child-directed speech samples. The update procedure shifts belief, represented here as probability, between competing hypotheses. All the modeled learners presented use online update procedures, meaning that they extract information from the data as the data come in. This is in contrast to learners (such as ideal/rational learners) that store all the data encountered to analyze together at some future point (e.g.,

⁵ It should be noted that most modeled learners converged before the “six months” (as measured in words encountered) were up. This suggests that even if a longer period were given for acquisition to occur (e.g., twelve months or eighteen months), the results reported here would not change.

Perfors, Tenenbaum & Regier 2006; Goldwater, Griffiths, & Johnson 2007; Foraker, Regier, Kheterpal, Perfors, & Tenenbaum 2007, 2009). Often such modeled learners are addressing the learnability of the information from the available data, without the constraints on processing that acquisition would require. Those studies complement studies that use incremental models (Gibson & Wexler 1994, Niyogi & Berwick 1996, Sakas & Fodor 2001, Yang 2002, Sakas 2003, Fodor & Sakas 2004, Gambell & Yang 2006, Pearl & Weinberg 2007, Vallabha, McClelland, Pons, Werker, & Amano 2007, Pearl 2008, Pearl & Lidz 2009), and these latter studies are more likely to use algorithms that are closer to the procedures children use to acquire language. Specifically, from the consideration of psychological plausibility, it is unlikely that children (or adults) have large enough memory capacity to store every utterance ever heard in all its detail. Instead, it seems far more likely that children process the data into smaller chunks, perhaps one or at most a few data points at a time, updating their hypotheses about the underlying system as they go.

4.3. Unbiased models

The basic hypothesis space for each of the unbiased models considered is the set of 156 viable grammars, comprised of the five main and four sub-parameters in the metrical phonology system. For each parameter, there are two competing values (e.g., QS vs. QI for quantity sensitivity). The learner initially associates a probability of 0.5 with each, representing no bias for either parameter value. This probability is then altered, based on the data encountered.

A given data point contains two types of information: the syllable rime structure and the stress contour. For example, the word *cucumber* has the syllable rime structure ‘VV VC VC’ and the stress contour ‘stressed stressed unstressed’. For each data point, the model generates a grammar based on the current probabilities associated with all parameter values, following the algorithm in Yang (2002). For instance, when generating the quantity sensitivity value, the modeled learner uses the probabilities associated with QI and QS. Suppose they are 0.40 and 0.60 respectively; then, the model will use the QI value with 40% probability and the QS value with 60% probability. If the learner uses the QS value, the sub-parameter QS-VC-H vs. QS-VC-L is then chosen based on the associated probabilities. This generation process continues until all parameter values have been selected. Using the probabilistically generated grammar, the model then constructs a stress contour for the word, given its syllable rime structure. If the generated stress contour matches the observed stress contour, all parameter values in that grammar are rewarded (11a); if the generated stress contour does not match, all parameter values in that grammar are punished (11b). Note that the learner does not attempt to assign credit or blame to a particular parameter value within the grammar. Instead, all participating values are rewarded or punished together, based on the grammar’s ability to match the observed stress contour. The learner then moves on to the next data point.

(11) Observed Stress Contour: *cucumber*

(a) grammar selected: QI, Em-None, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Left

generated stress contour:

syllable class (S) (S S)

syllables cu cum ber
match: reward all

(b) grammar selected: QI, Em-None, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Rt

generated stress contour:

syllable class (S) (S S)
syllables cu cum ber
mismatch: punish all

When the probability for one parameter value approaches 0.0 or 1.0, the learner sets that parameter to the appropriate value. For example, if the threshold was 0.2/0.8 and Em-Some's probability reached 0.8, the learner would set the extrametricality parameter to Em-Some by giving that parameter value a probability of 1.0 (while Em-None would be set to a probability of 0.0). The grammar generated for subsequent data points would then always contain the value Em-Some, since its probability is 1.0. All simulations used a 0.2/0.8 threshold, based on estimates of the thresholds children are able to generalize at (Gómez & Lakusta 2004, Hudson Kam & Newport 2005).⁶ Ideally, after a reasonable

⁶ Specifically, Hudson Kam & Newport (2005) show that 6-year-old children do not reliably extract a generalization from noisy data when the probability of the generalization occurring in the data is 0.60. This suggests that the threshold for generalizing is higher than this. Gómez & Lakusta (2004) demonstrate that 12-month-old children seem able to make a generalization from noisy data when the probability of the generalization occurring in the data is 0.83 (but not when the probability is 0.67). This suggests that the threshold is above 0.67 but may be lower than 0.83. The threshold of 0.8 for these simulations was chosen

number of English data points, the learner will set the correct values for the English grammar.

The unbiased learners considered here vary with respect to how they implement the reward/punishment component of the update procedure. One learner type is based on the Naïve Parameter Learner (NPLearner) described in Yang (2002), which uses the Linear reward-penalty scheme (Bush & Mosteller 1951), as shown in (12). The update equation involves a parameter γ that determines how liberal the learner is. The larger γ is, the more probability the learner shifts for a single data point.

(12) Linear Reward-Penalty Scheme

p_v = previous probability of parameter value (e.g., QI)

p_o = previous probability of opposing parameter value (e.g., QS)

(a) generated stress contour matches observed stress contour (reward)⁷

$$p_{v_{new}} = p_v + \gamma(1 - p_v)$$

$$p_{o_{new}} = 1 - p_{v_{new}}$$

(b) generated stress contour does not match observed stress contour (punish)

$$p_{v_{new}} = (1 - \gamma)p_v$$

$$p_{o_{new}} = 1 - p_{v_{new}}$$

as a value within this range. Varying this threshold value between 0.67 and 0.83 did not qualitatively change the results found.

⁷ Though there is actually a separate formula for calculating $p_{o_{new}}$, we can calculate it this way since there are only two values for any parameter. Note also that multiplying the terms out will show a more intuitive notion of reward and punishment: both the reward and punishment of p_v involve the subtraction of the quantity $\gamma * p_v$ from the original probability, and only the reward involves the addition of the quantity γ .

As an example, suppose we consider the probabilities of QI and QS for the quantity sensitivity parameter. Initially, they are both 0.5. For the first data point, suppose QI is chosen to be part of the grammar (using the probabilistic grammar generation process described at the beginning of this section) and that grammar fails to generate the observed stress contour. The QI value (and all other participating values) are punished. Suppose γ is 0.01. The new value of QI would be $(1-0.01)*0.5 = 0.495$ and the new value of QS would be $1-0.495 = 0.505$.

The second learner type is a Bayesian learning variant (BayesLearner) that uses Bayes' rule to update parameter value probability. Since there are only two parameter values per parameter, the learner uses the beta distribution to calculate the probability a binomial distribution should be centered at in order to account for the observed data (Chew 1971). The update equation involves two statistical parameters, α and β (see (13)). Setting both of these values to 0.5 initially biases the model to favor neither parameter value, and also to prefer the probabilities reflected by the observed data. This is because these values represent prior beliefs the learner has, specifically that the learner imagines (before ever observing any data) that 0.5 of 1 data point supports one parameter value over the other.⁸ Since these values are so small, they represent a very weak initial

⁸ Obviously, learners cannot really observe fractions of data points. However, prior beliefs represent probabilities, so this can also be thought of as a learner having 50% confidence that 1 data point supports one parameter value over the other. We could encode this same idea by having $\alpha = \beta = 1$, so that the prior belief is that 1 out of 2 data points support one parameter value over the other. However, this means the initial bias for a parameter probability of 0.5 is slightly stronger, and takes more observed data to overcome.

bias, and the observed data will soon overshadow this bias. This means the learner has no initial preference for a parameter's value, and is strongly driven by the observed data. If a parameter value participates in a grammar that generates a matching stress contour, the number of successes for that parameter value is incremented by 1. If a parameter value participates in a grammar that does not, the number of successes is left alone. Either way, the total data seen is incremented by 1 if the parameter value was part of the grammar used to generate the stress contour. The probabilities for opposing parameter values are then calculated and all probabilities are normalized so they sum to 1. So, for each parameter value, the model tracks (a) the current probability, (b) the number of matching stress contours that parameter value has been involved in generating, and (c) the total number of stress contours that parameter value has been involved in generating.

(13) BayesLearner update equation

p_v = previous probability of parameter value (e.g., QI)

p_o = previous probability of opposing parameter value (e.g., QS)

$$p_{v_{new}} = \frac{\alpha + 1 + \text{successes}}{\alpha + \beta + 2 + \text{total data seen}}$$

$$p_{v_{new}, \text{normalized}} = \frac{p_{v_{new}}}{p_{v_{new}} + p_o}$$

$$p_{o_{new}, \text{normalized}} = \frac{p_o}{p_{v_{new}} + p_o}$$

As an example, suppose we consider the same scenario as before: the probabilities of QI and QS for the quantity sensitivity parameter. Initially, they are both 0.5. For the first data point, suppose QI is chosen to be part of the grammar and that grammar fails to

generate the observed stress contour. The QI value (and all other participating values) are punished. The non-normalized probability for the QI value is $(0.5+1+0)/(0.5+0.5+2+1) = 0.375$. The non-normalized probability for the QS value has not changed from 0.5 since it was not used for this data point. The normalized probability of QI is then $0.375/(0.375 + 0.5) = 0.429$ while the normalized probability of QS is then 0.571.

One property of these learners is that neither is very noise-tolerant, since the probabilities are updated for each data point encountered. Given the noisy English data and the complex system with interacting parameters, this may not be a desirable property. Yang (2002) advocates a method called *batch-learning* for smoothing the acquisition trajectory when the system to be acquired involves multiple parameters, such as the metrical phonology system here. Unlike the standard usage of the term batch-learning, this method does not require the learner to analyze larger quantities of data simultaneously. Instead, the learner simply keeps a count of how many successes (matches) or failures (mismatches) a parameter has had in a row. If the parameter has succeeded or failed a certain number of times in a row, only then does the learner invoke the update function. Thus, this method is compatible with an incremental learning procedure that extracts information from data as they come in. In addition, this method allows the learner to be more robust in the face of noisy data, as a string of successes/failures is less likely to result unless that parameter value really is succeeding/failing on the majority of the data. In order to distinguish this method from the standard usage of batch-learning, we will refer to it as *count-learning* hereafter.

The count size c regulates how often a parameter value is rewarded/punished. Every time the parameter value is part of a grammar that generates a matching stress contour, that parameter value's counter is incremented; every time the parameter value is part of a grammar that generates a mismatching stress contour, that parameter value's counter is decremented. If the counter reaches c , the parameter value is rewarded; if the counter reaches $-c$, the parameter value is punished. Afterwards, the counter is reset to 0. Applying count-learning to the learner types already discussed is straightforward. A count NPLearner will reward/punish a parameter value if the counter reaches $\pm c$. A count BayesLearner only updates if the counter reaches $\pm c$: specifically, if the counter is $+c$, *successes* is incremented by 1 and *total data seen* is incremented by 1; if the counter is $-c$, only *total data seen* is incremented by 1.

We illustrate the count version of each learner type below with an example. Suppose we again consider the parameter values QI and QS for the quantity sensitive parameter. Suppose that c is 5. Initially, QI and QS both have probability 0.5, and their counters are both 0. For the first data point, suppose QI is chosen to be part of the grammar and that grammar fails to generate the observed stress contour. The counter for QI is now -1. For the next three data points, suppose QS is chosen for the grammar and those grammars succeed at generating the observed stress contour. The counter for QS is +3 and the counter for QI is -1. Suppose the next two data points use QI and those grammars succeed: QS's counter is still +3, but QI's counter is now +1. Suppose then that the next six data points use QI and those grammars fail: QS's counter is still +3, but QI's counter is now -5, which is the count limit c . The QI value is then punished using the appropriate update equation for the learner. If the NPLearner uses a γ of 0.01, the new probability of

QI is 0.495 and the new probability of QS is 0.505. If the BayesLearner learner is used, the new probability of QI is 0.429 and the new probability of QS is 0.571. The counter for QI is then reset to 0.

The count-learner's robustness to noise can be seen from the previous example – instead of updating for each of the twelve individual data points (punishing QI once, rewarding QS three times, rewarding QI two times, and then punishing QI six times), the learner only punishes QI once. Importantly, this is only after the QI value has been involved in a string of failures, and so is more likely to really be failing to be compatible with the data.

4.4. Processing the input

Since a data point consists of a single word at a time, the learners here included the assumption that children can successfully identify words in fluent speech by the time they are acquiring the metrical phonology system. This does not seem unreasonable as word segmentation research by Jusczyk and colleagues suggests that children as young as seven months can identify some words in fluent speech successfully (Jusczyk & Aslin 1995, Jusczyk, Houston, & Newsome 1999), so this process should be operational by the time children are acquiring a generative system of stress. In addition, each data point was pre-divided into syllables, with individual syllables identified by rime as type VV(C), VC, or V. Thus, the learners also included the assumption that children can successfully syllabify words and are sensitive to the rime structure. This also does not seem unreasonable as Jusczyk and colleagues have suggested that young infants are sensitive to syllables and properties of syllable structure (Jusczyk, Goodman, & Baumann 1999,

Turk, Jusczyk, & Gerken 1995), so this process would again likely be operational by the time acquisition of the stress system begins. Thirdly, the learners did not call the update procedure if a monosyllabic word was encountered, as monosyllabic words do not have a stress contour (i.e., a sequence of syllables that are stressed/unstressed relative to each other within a given word). Instead, monosyllabic words were ignored. This can be viewed as a learner assumption that the generative system is used for defining the contour over multisyllabic sequences that will have relatively contrasting stress among the syllables, and is not used when only a single syllable is present. Under this view, monosyllabic words are not informative.⁹ Fourthly, the learners did not set any sub-parameters before the corresponding main parameter was set. For example, the quantity sensitivity sub-parameter QS-VC-L vs. QS-VC-H would not be set before setting the main quantity sensitivity parameter QS. So, until QS was set, no data impacted the probabilities of QS-VC-L and QS-VC-H. This assumes that children will only consider information about a sub-parameter if it is necessary to do so to acquire their particular language's grammar; otherwise, they will not bother tracking the success rate for that sub-parameter.

⁹ Also, it turned out that simulations with models that processed monosyllabic words never converged on the English grammar due to the extrametricality parameter. Since the majority of monosyllabic words are stressed, the English property of having extrametricality on the rightmost syllable (Em-Some, Em-Right) was punished by these data, as the rightmost syllable was the only syllable in the word. Since that syllable is stressed, it cannot be extrametrical. So, a stressed monosyllabic word is incompatible with an analysis that requires Em-Some. This suggests that a learner trying to acquire the English grammar for this parametric system must assume monosyllabic words are not informative.

4.5. Learner parameters and simulations

The four learners – NPLearner, BayesLearner, Count NPLearner, and Count BayesLearner – were run on the input set, which was probabilistically generated from the English child-directed speech distributions. So, while the probability distribution of the data each learner used was the same, the exact words and order in which these words were encountered varied for each run of a modeled learner, creating a randomized data set with similar distributional properties to the English dataset. The NPLearner and Count NPLearner were run with learning parameter $\gamma = 0.001, 0.0025, 0.01, \text{ and } 0.025$. The Count NPLearner and Count BayesLearner were run with count parameter $c = 2, 5, 7, 10, 15, \text{ and } 20$. Each learner variant was run 1000 times. The desired output behavior was to converge on the English grammar within the acquisition period, as defined by the number of data points an average child would encounter in six months (1,666,667).

5. Results and discussion

Table 1 shows the percentage of the trials each learner converged on the English grammar. When multiple parameter values are used for a learner (e.g., $c = 2, 5, 7, 10, 15, \text{ or } 20$ for the counting variants), the average percent convergence is given.

Unbiased Learner	% English Convergence (.01 = .01%)
NPLearner $\gamma = 0.001, 0.0025, 0.01, \text{ or } 0.025$	0.000

BayesLearner	0.000
Count NPLearner $\gamma = 0.001, 0.0025, 0.01, \text{ or } 0.025$ $c = 2, 5, 7, 10, 15, \text{ or } 20$	0.033
Count BayesLearner $c = 2, 5, 7, 10, 15, \text{ or } 20$	0.000

Table 1. Unbiased learner results.

The most striking aspect of these results is the extreme rarity with which these unbiased learners converge on the English grammar. Only the Count NPLearner ever manages to do it, and then only for about one out of every 3000 trials.

How do we interpret this lack of convergence for unbiased probabilistic learners? Recall that the biased learner from Pearl (2008, 2009) guaranteed convergence so long as the child learned only from unambiguous data (Fodor 1998a, Dresher 1999, Lightfoot 1999) and set the parameters in a particular order (See Appendix A3 for details of how this works). Thus, convergence in a small fraction of the trials here certainly does not look like the best we can do. If we look closer at the modeling results obtained here, we can see what kind of errors the unbiased learners are making. It seems in general that these learners will converge on grammars that have several parameter values in common with the English grammar – but crucially are different on at least one value. In (14), we see several example grammars of this kind, with incorrect values in *italics*. Unfortunately for our learners, having even one value incorrect means that the grammar is not correct,

and will produce contours different from the English grammar (recall example (9) where changing only one parameter value severely alters the generated stress contour).

(14) Examples of incorrect grammars selected by unbiased learners

(a) *QI*, *Em-Some*, *Em-Right*, *Ft-Dir-Left*, *Unb*, *Ft-Hd-Left*

(b) *QI*, *Em-None*, *Ft-Dir-Rt*, *Unb*, *Ft-Hd-Left*

(c) *QS*, *QS-VC-H*, *Em-Some*, *Em-Right*, *Ft-Dir-Rt*, *B*, *B-2*, *B-Mor*, *Ft-Hd-Left*

(d) *QS*, *QS-VC-L*, *Em-Some*, *Em-Right*, *Ft-Dir-Rt*, *Unb*, *Ft-Hd-Rt*

5.1. The problem for unbiased learners

Obviously, this exceptionally poor performance was not the behavior we were looking for from these unbiased learners. The question then is whether the problem is with these modeled learners in particular, or if there is some underlying issue that will cause all unbiased probabilistic models to fail. If the problem is with these learners, then we simply need to try better learners. For instance, perhaps if we did not constrain our learners so much, they would perform better. This would be because, by constraining our learners, we have somehow hindered them from finding the optimal grammars in the hypothesis space. However, if the problem is somehow inherent to the acquisition task as defined, then no unbiased probabilistic model can be successful – constrained or not.

Let us examine the acquisition task in more detail. The hypothesis space contains 156 grammars: the grammar proposed for English and 155 others. We know that the English grammar is not compatible with all the available English data (as indeed none of the grammars are), but how compatible is it? It turns out that the English grammar

generates stress contours that are compatible with 73.0% of the observable data tokens (where every instance of a word is counted) and 62.1% by types (where frequency is factored out and a lexicon item is only counted once no matter how often it occurs). So this English grammar covers a large portion of the data, even if it requires a non-trivial number of words to be viewed as exceptions to the system (see Appendix A2 for more discussion of words that are exceptions for this English grammar).

However, let us now look at the grammars our constrained learners are choosing, given the data. The average compatibility of these grammars is 73.6% by tokens and 63.3% by types, which is slightly higher than that of the English grammar. Some of the grammars commonly chosen are listed below in Table 2, along with their token and type compatibility scores:

Grammar	Token Compatibility	Type Compatibility
<i>QS-VC-L</i> , Em-Some, Em-Right, <i>Ft-Dir-Left</i> , <i>Unb</i> , Ft-Hd-Left	73.4%	64.6%
<i>QS-VC-L</i> , Em-Some, Em-Right, Ft-Dir-Rt, <i>Unb</i> , Ft-Hd-Left	73.4%	63.7%
<i>QI</i> , <i>Em-None</i> , Ft-Dir-Rt, <i>Unb</i> , Ft-Hd-Left	73.6%	63.3%
<i>QI</i> , <i>Em-None</i> , <i>Ft-Dir-Left</i> , <i>Unb</i> , Ft-Hd-Left	73.6%	63.3%

Table 2. Non-English grammars commonly selected by unbiased probabilistic learners.

Non-English parameter values are *italicized*.

The grammar with the highest compatibility in the hypothesis space - which differs from the English grammar in quantity sensitivity, extrametricality, and foot directionality (QI, Em-None, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left) - has scores of 76.5% by tokens and 70.3% by types, which is again not all that much higher than the English grammar's compatibility. But the simple fact remains: the grammars the constrained learners are choosing *are* the better grammars. The problem is that the English grammar is not one of those more optimal grammars.

How far from optimal is it? If we rank the competing grammars by their compatibility with the English data set, it turns out that there are 51 other grammars that are more compatible with the data tokens than the English grammar. If we make the comparison to the data types, the English grammar is less compatible than 55 other grammars. That is, the English grammar is barely in the top third of the hypothesis space, when ranked by compatibility with the child-directed speech data. Given this, it is not surprising that our unbiased probabilistic learners rarely chose it – unbiased probabilistic learners are geared to identify the more optimal hypotheses in the hypothesis space, and it turned out that the English grammar was not one of those hypotheses.

5.2. The real culprits

The behavior of the unbiased probabilistic learners does not accord with the behavior we expected to see in children, given our definition of the acquisition task. Since the problem is not with the unbiased probabilistic learners, which are performing just as they should to identify optimal grammars, then where does the problem lie?

5.2.1. The initial target state

One option is that we have modeled acquisition very well with these unbiased learners, but the target state for children is not what we thought it was. Specifically, recall that this parametric knowledge representation (as with many other theories of knowledge representation) was constructed based on analysis of adult usage. It is possible that, while these values are correct for English stress, the English grammar is only optimal when the full range of word forms in English are available. Thus, the child would only converge on the English grammar after more adult-like speech has been encountered.

To test this, we can examine a corpus of adult-directed speech and assess the English grammar's compatibility with those data, as well as its overall ranking in comparison to the other grammars in the hypothesis space. The North American English CALLFRIEND corpus (Canavan & Zipperlen 1996) contains transcripts of phone calls between adult English speakers in the Northern United States, as recorded by the Linguistic Data Consortium. It includes 82,487 data tokens and 4,719 data types, with non-monosyllabic words comprising 14,235 data tokens and 2,851 data types (see Appendix A1 for more details of this corpus). The English grammar defined here is compatible with 63.7% of the data tokens and 52.1% of the data type – even worse than this same grammar's compatibility with the child-directed speech data! However, there are fewer grammars that can do better - only 33 other better grammars by data token compatibility and 35 better by data type compatibility, with the most compatible grammar (QI, Em-Some, Em-Right, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Left) parametrically very similar to the English grammar (only the quantity sensitivity value differs) and

having a data token compatibility of 66.6% and a data type compatibility of 56.3%.

Thus, the English grammar is more optimal in the hypothesis space when adult-directed speech are considered, and so more likely to be chosen by an unbiased probabilistic learner.

However, the lower overall compatibility of the English grammar with the adult-directed speech data highlights another important consideration for the acquisition task. As mentioned in the initial discussion of this parametric system, this system does not include interactions with the morphology system, which are important for completely characterizing the adult knowledge state for English (Chomsky & Halle 1968, Kiparsky 1979, Hayes 1982).¹⁰ Yet we supposed that children initially would not realize the connection between the two systems, and so would not be able to use this knowledge to account for more data. So, perhaps children's initial target state should not be the English grammar values as currently defined; instead, their initial target state is for a set of parameter values that are optimized for a system that lacks interaction with morphology. Given child-directed speech data, they indeed would find the grammars our learners have found here. We should then expect to observe an extended period of time where they believe a non-English parameter value is correct, and this should last until the morphology system comes online and they realize the potential interactions between morphology and stress (perhaps around age 3, as Brown (1973) shows robust usage of morphological markers such as progressive and plural inflections in child speech by 3;1).

¹⁰ We could imagine that models containing some morphological knowledge might be able to converge on the English grammar, given adult-directed speech or perhaps even given child-directed speech. This has been left for future research.

Kehoe (1998) conducted an elicitation task with children on English words and novel words following English stress patterns, and found by studying the errors in children's productions that 28-month-olds seemed to be experimenting with aspects of the metrical phonology system such as quantity sensitivity and extrametricality while 34-month-olds were mostly able to imitate the correct stress pattern. She took this to mean that the 34-month-olds had acquired the English system, but it is possible that in fact children were using a non-English system capable of producing the stress contours they were tested on. For example, one novel word children were tested on was “tanema” (rhyming with “Panama”: /tænəmə/), which had the stress pattern of “stressed unstressed unstressed”. Around 41% of the time, 34-month-old children imitated the correct stress contour, while around 52% of the time, these children produced 2-syllable truncations with stress on the first syllable (e.g., “tama”). (The rest of the time, these children produced different random errors.) We can see in (15) below that this behavior could be compatible both the English grammar and all of the grammars listed in table 2 that were commonly chosen by our unbiased probabilistic learners.

(15) “tanema” analyses (syllable rime structure is V V V)

(a) the English grammar & two commonly chosen grammars produce the same analysis:

QS, QS-VC-H, Em-Some, Em-Right, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Left

QS, QS-VC-L, Em-Some, Em-Right, Ft-Dir-Left, Unb, Ft-Hd-Left

QS, QS-VC-L, Em-Some, Em-Right, Ft-Dir-Rt, Unb, Ft-Hd-Left

full word

2-syllable truncation

(<u>L</u> L) <L>	(<u>L</u>) <L>
V V V	V V
<u>ta</u> ne ma	<u>ta</u> ma

(b) the other two commonly chosen grammars produce the same stress contour:

QI, Em-None, Ft-Dir-Left, Unb, Ft-Hd-Left

QI, Em-None, Ft-Dir-Rt, Unb, Ft-Hd-Left

full word	2-syllable truncation
(<u>S</u> S S)	(<u>S</u> S)
V V V	V V
<u>ta</u> ne ma	<u>ta</u> ma

To gauge what English children’s initial target state is (around say, 3 years, before the morphology system is fully online), we might wish to conduct similar elicitation style experiments with an eye towards forms that would more specifically single out one grammar from another.¹¹ For example, we might expect a form like “toynema” (/tɔjnəmə/) to differentiate between the English grammar and one of the grammars in (15a), as the English grammar would still assign this the same stress contour as before while that non-English grammar selected by our unbiased probabilistic learners would

¹¹ It may also be possible to gauge children’s knowledge with some kind of task that does not require as much effort from them, since an imitation or elicitation task may have other performance factors associated with it. To this end, it might be useful to have children choose which pronunciation they prefer for a given novel word.

produce different stress contours in some cases, as shown in (16).

(16) Stress contours for “toynema” (syllable rime structure is VV V V)

(a) English grammar

QS, QS-VC-H, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Left

full word	2-syllable truncation
(<u>H</u> L) <L>	(<u>H</u>) <L>
VV V V	VV V
<u>toy</u> ne ma	<u>toy</u> ma

(b) non-English grammar

QS, QS-VC-L, Em-Some, Em-Right, Ft-Dir-Rt, Unb, Ft-Hd-Left

full word	2-syllable truncation
(<u>H</u>) (<u>L</u>) <L>	(<u>H</u>) <L>
VV V V	VV V
<u>toy</u> <u>ne</u> ma	<u>toy</u> ma

(different contour)

In summary, given our incomplete knowledge of young children’s metrical competence, it is possible that the unbiased probabilistic learners examined here actually performed quite well when assessed in terms of the parametric system under consideration. Specifically, the target grammar for young children without morphological knowledge may not be the adult grammar proposed by Dresher (1999).

We have suggested one kind of experiment that may more precisely identify young children's metrical phonology knowledge, so that this question can be decided.

5.2.2. Biased learning

Another possibility is that children do indeed reach the adult English grammar defined here, and we will find that their productions and knowledge are consistent with only the English grammar. However, to acquire this parametric system, they must use some kind of bias to help them along. That is, in order to use the probability distributions in the data effectively, children must not be unbiased learners.

A bias that was found to be successful in previous modeling work for this same case study was a selective learning bias that altered the learner's intake (Pearl 2008, 2009). In particular, the modeled learner only learned from the subset of the available input viewed as maximally informative: unambiguous data (Fodor 1998a, Dresher 1999, Lightfoot 1999, Pearl & Weinberg 2007). While learning only from maximally informative data has intuitive appeal, it is not without its difficulties. Specifically, data are often ambiguous, especially in systems involving multiple interacting parameters, such as metrical phonology. So, unambiguous data would comprise only a small subset of the available input. Moreover, it is not necessarily straightforward to identify unambiguous data, though there are various proposals for how children might be able to do this such as by looking for cues in the observable data (Dresher 1999) or by parsing the data with various parameter value combinations (Fodor 1998b, Sakas & Fodor 2001). See Appendix A3 for more discussion of unambiguous data for the parameters discussed in this case study, and the proposals for identifying unambiguous data in the input.

Notably, Pearl (2008, 2009) found that a general class of probabilistic learners was successful as long as they employed this selective learning bias and obeyed certain parameter-setting order constraints, such that some parameters were learned before other parameters (see Appendix 3 for details). The reason why this bias works is because the unambiguous data favor the English parameter values when the parameters are acquired in particular orders. For example, the unambiguous data for the extrametricality value Em-None may outnumber the unambiguous data for Em-Some before the learner realizes the grammar is quantity sensitive (QS). However, once the grammar is known to use QS, the learner may then alter her view about which data are informative (e.g., if the rightmost syllable is Heavy but unstressed, this is a signal that this syllable is extrametrical). It may then turn out that, with this new knowledge, the data perceived as unambiguous for Em-Some now outnumber the data perceived as unambiguous for Em-None. A probabilistic learner should then choose Em-Some after observing a sufficient quantity of data.

If the probabilistic learner learns the parameters in one of the viable parameter-setting orders, the English grammar is the grammar selected, since the parameter values comprising the English grammar are favored by the unambiguous data distributions. Success then rests on the child having knowledge of these viable parameter-setting orders. Depending on the method used to identify unambiguous data, the knowledge of the appropriate orders may be derivable from either the data themselves or other learning biases the child has (see Pearl (2007, 2009) for discussion).

Interestingly, simulations have suggested that it is not the parameter-setting order alone that causes the English grammar to be chosen – when parameters are set in similar

orders by the unbiased learners described here, there is still no reliable convergence on the English grammar. Moreover, the few times these learners do converge on the English grammar, there is no commonality in their parameter-setting order. The only other potential cause of the desired acquisition behavior from the biased learners is their restriction to unambiguous data. A reasonable question is why the unambiguous data do not exert their influence sufficiently within the larger dataset of the input. That is, since the unambiguous data are present and appear in the correct distributions to lead the child to the English grammar (subject to certain parameter-setting order constraints), why don't they do so even if other unhelpful data are present? The answer may have to do with the quantity of unambiguous data. For this case study, the unambiguous data are a small minority of the available data (at most around 5%). The ambiguous data, by definition, are compatible with competing grammars. So, it is likely that the helpful bias the unambiguous data provide is washed away in the wake of the ambiguous data that must be processed, since probabilistic learners extract some information from ambiguous data as well, and that information may lead the learners astray.

The implication of this success with biased learners is that the acquisition task as currently defined (i.e., learn the adult English grammar from child-directed English data, without morphological knowledge) is not impossible for probabilistic learners; it just may be impossible for unbiased probabilistic learners¹². If we discover that children do indeed select the English grammar initially, even prior to morphological knowledge, then this suggests that children must have some biases to guide their acquisition if they are

¹² Note that this result specifically applies to learners attempting to learn using this parametric knowledge representation. Unbiased probabilistic learning may succeed when learners are using alternative knowledge representations, which then would support those knowledge representations over the knowledge representation considered here (see discussion in section 6).

representing their knowledge of metrical phonology the way this parametric system believes them to do.

6. Implications for acquisition and theories of knowledge representation

What we have learned from the computational models presented here is what is necessary for children who use probabilistic learning to succeed at this acquisition task, as it is currently defined. Specifically, for children to select the adult English grammar in this parametric system, given child-directed speech data and no knowledge of the connection between metrical phonology and the English morphology system, children cannot be unbiased. This is an argument from acquisition for a bias in children, if they are to use this parametric system to learn the adult English grammar defined in this system.

Now, as discussed in the previous section, we may discover through experimental research that children's initial target state is not the adult English grammar. If so, we can then see if unbiased learners are selecting the grammars children are selecting, given the data. If we find that unbiased learners are *still* not selecting the same grammars as children, we again have a reason to believe children require a bias to guide their acquisition of this parametric system. If, however, we find that both unbiased learners and children are indeed selecting the same grammars, then we have even stronger support for this parametric knowledge representation – nothing beyond a sensitivity to data distributions is required to learn as children seem to learn. Note, however, that support

for this knowledge representation does not necessarily rule out other knowledge representations that might be similarly compatible with children's metrical competence.¹³

To that end, the methodology used in this paper to examine this parametric knowledge representation is something that can be applied to many different theories of knowledge representation (many of which may also be motivated primarily by their ability to account for limited cross-linguistic variation). For each theory, we can explore an argument from acquisition by seeing what kinds of probabilistic learners – unbiased or biased – can acquire the correct grammar from child-directed speech data, assuming that theory's knowledge representation. We can then compare theories of knowledge representation, based on their acquisition performance. For instance, perhaps two theories allow the correct grammar to be selected based on child-directed speech data, but one requires probabilistic learners to have some bias while the other does not. This could be taken as support for the theory that does not require biased learners, as it assumes less knowledge in the learner. As another example, suppose one theory allows the correct grammar to be selected only by biased learners, while another theory does not allow the correct grammar to be selected at all. This second theory might be dispreferred since it is “unacquirable” from child-directed speech data, even though the first theory still requires additional knowledge (in the form of a bias).

Using this framework, we can explore the acquirability of other knowledge representations proposed within a generative framework. For instance, Hayes (1995) specifies a parametric system that includes a more restricted inventory of metrical foot

¹³ In addition, support for this generative knowledge representation does not rule out alternative non-generative approaches to acquisition that also may be compatible with children's observable knowledge (e.g., the exemplar-based approach of Daelamans *et al.* (1994)). Notably, however, these alternative approaches do not necessarily have anything to say about the constrained variation seen cross-linguistically, which is often one of the motivations for generative systems (Hayes 1995).

types than what the system here assumed, combining aspects of quantity sensitivity, boundedness, and foot headedness. It is possible this more restricted system allows the correct grammar to be acquired from child-directed speech without the probabilistic learners having any additional knowledge. We can also explore optimality theoretic systems, where constraints must be ranked rather than parameters set (Tesar & Smolensky 2000). It is again possible that by using a different knowledge representation, and so changing what the child is trying to learn, the English child-directed speech make the English grammar (however it may be defined) much more favorable for unbiased learners.

At the same time as we compare knowledge representations, we can also gain knowledge about the acquisition process for generative knowledge representations. Suppose we discover that no matter what knowledge representation we use, unbiased learners still cannot succeed on English metrical phonology. This tells us some kind of bias is required for acquiring a generative system, an idea noted by several researchers for other case studies in acquisition (e.g. English anaphoric *one*: Regier & Gahl 2004, Foraker *et al.* 2009, Pearl & Lidz 2009; structure-dependence of syntactic rules: Perfors, Tenenbaum, & Regier 2006). We can then explore what biases may allow probabilistic learners to succeed. The unambiguous data bias discussed in the previous section is one bias that is successful for a particular implementation of a parametric system, but there may well be others that accomplish the same thing. For instance, a child may have a bias to learn only from data that appear to be systematic or productive (Yang 2005). The exact nature of the necessary bias can be investigated through computational modeling studies, such as Perfors, Tenenbaum, & Regier (2006) and Foraker *et al.* (2009), which

use a simplicity bias on the hypothesis space, and Regier & Gahl (2004), Pearl (2008, 2009), and Pearl & Lidz (2009), which use a subset bias on the hypothesis space and a data intake bias. Of particular interest is whether the necessary bias is likely to be domain-specific (Regier & Gahl 2004, Pearl 2008, 2009, Pearl & Lidz 2009) or domain-general (Regier & Gahl 2004, Perfors, Tenenbaum, & Regier 2006, Foraker *et al.* 2009, Pearl & Lidz 2009). For instance, while the bias to learn only from unambiguous data may be domain-general in nature, the identification of unambiguous data may be domain-specific in nature (see Pearl (2007) for discussion on this point).

7. Conclusion

We have examined a parametric knowledge representation for metrical phonology from the perspective of acquisition, using computationally modeled learners that attempted to acquire the adult English grammar based on realistic distributions of English child-directed speech. We discovered that solving this acquisition task using this knowledge representation requires the learner to have something more than the ability to leverage probabilistic information in the data. In the broader picture, we have presented a computational modeling framework that allows us to make an argument from acquisition for (or against) a particular knowledge representation, based on the acquirability of the correct grammar for that knowledge representation. At the same time, we can specify what is required to make the correct grammar acquirable, thereby describing more precisely the acquisition process for a child using a particular knowledge representation. Key to this approach is that the modeled learners learn from realistic data and consider psychological plausibility in their algorithms; when they do this, modeled learners make

more powerful arguments from acquisition since they are more closely approximating the child's acquisition process. This approach can allow us to understand how children solve the acquisition problems that they do, and what knowledge they are using while they do it.

Acknowledgements:

Many thanks to Amy Weinberg, Bill Idsardi, Jeffrey Lidz, Charles Yang, Roger Levy, Jon Sprouse, Ivano Caponigro, Diogo Almeida, Heather Goad, Diane Lillo-Martin, four anonymous reviewers, and the audiences at GALANA 2008, the UCSD Linguistics Department, the UC Irvine Artificial Intelligence and Machine Learning Group, the UCLA Linguistics Department, the USC Linguistics Department, and the UMCP Linguistics Department. This work has been supported by NSF Grant BCS-0843896.

References

- Akhtar, Nameera, Maureen Callanan, Geoffrey Pullum, & Barbara Scholz. 2004. Learning antecedents for anaphoric *one*. *Cognition* 93. 141-145.
- Archibald, John. 1992. Adult abilities in L2 speech: evidence from stress. In Jonathan Leather & Allan James (eds.), *New Sounds 92: Proceedings of the 1992 Amsterdam Symposium on the Acquisition of Second Language Speech*, 1-17. Amsterdam: University of Amsterdam Press.
- Bernstein Ratner, Nan. 1984. Patterns of vowel Modification in motherese. *Journal of Child Language* 11. 557-578.
- Brent, Micahel and Jeffrey Siskind. 2001. The Role of Exposure to Isolated Words in

- Early Vocabulary Development. *Cognition* 81/82. 33–44.
- Brown, Roger. 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.
- Bush, Robert & Frederick Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 58. 313-323.
- Canavan, Alexandra and George Zipperlen. 1996. CALLFRIEND American English-Non-Southern Dialect. Linguistic Data Consortium: Philadelphia, PA.
- Canavan, Alexandra, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech. Linguistic Data Consortium: Philadelphia, PA.
- Chew, Victor. 1971. Point Estimation of the Parameter of the Binomial Distribution. *American Statistician* 25(5), 47-50.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
- Clark, Robin. 1992. The Selection of Syntactic Knowledge. *Language Acquisition* 2(2). 83-149.
- Clark, Robin. 1994. Kolmogorov complexity and the information content of parameters. IRCS Report 94-17. Institute for Research in Cognitive Science, University of Pennsylvania.
- Crain, Stephen & Paul Pietroski. 2002. Why language acquisition is a snap. *The Linguistic Review* 19. 163-183.
- Daelemans, Walter, Steven Gillis, and Gert Durieux. 1994. The Acquisition of Stress: A Data-Oriented Approach. *Association for Computational Linguistics* 20(3). 421-451.
- Dresher, Elan. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* 30. 27-67.

- Dresher, Elan & Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* 34. 137-195.
- Fodor, Janet Dean. 1998a. Unambiguous Triggers. *Linguistic Inquiry* 29(1). 1-36.
- Fodor, Janet Dean. 1998b. Parsing to Learn. *Journal of Psycholinguistic Research* 27(3). 339-374.
- Fodor, Janet Dean & William Sakas. 2004. Evaluating Models of Parameter Settings. *Proceedings of the 28th Annual Boston University Conference on Language Development*, 1-27. Somerville, MA: Cascadilla Press.
- Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors, & Joshua Tenenbaum. 2007. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. Nashville, TN.
- Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors, & Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Cognitive Science* 33(2), 287-300.
- Gambell, Timothy & Charles Yang. (2006). Word Segmentation: Quick but not dirty. Manuscript: Yale University.
- Gerken, LouAnn & Richard Aslin. 2005. Thirty years of research on infant speech perception: The legacy of Peter W. Jusczyk. *Language Learning and Development* 1, 5-21.
- Gibson, Edward. & Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25. 407-454.
- Gillis, Steven, Walter Daelemans, and Gert Durieux. 2000. Lazy Learning: A comparison of Natural and Machine Learning of Stress. In P. Broeder and J.M.J. Murre (Eds.),

- Models of Language Acquisition: inductive and deductive approaches*. Oxford University Press, 76-99.
- Goldberg, Adele. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.
- Goldwater, Sharon, Tom Griffiths, & Mark Johnson. 2007. Distributional cues to word segmentation: Context is important. *Proceedings of the 31st Boston University Conference on Language Development*, 239-250. Somerville, MA: Cascadilla Press.
- Gómez, Rebecca & Laura Lakusta. 2004. A first step in form-based category abstraction by 12-month-old infants. *Developmental Science* 7(5). 567-580.
- Halle, Morris & William Idsardi. 1995. General Properties of Stress and Metrical Structure. In Goldsmith, J. (ed.), *The Handbook of Phonological Theory*, 403-443. Cambridge, MA & Oxford: Blackwell Publishers.
- Halle, Morris & Jean-Roger Vergnaud. 1987. *An Essay on Stress*. Cambridge, MA: MIT Press.
- Hart, Betty & Todd Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: P.H. Brookes.
- Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.
- Heinz, Jeffrey. 2007. Learning Unbounded Stress Patterns via Local Inference. *Proceedings of the 37th Annual Meeting of the Northeast Linguistics Society (NELS 37)*.
- Hochberg, Judith. 1988. Learning Spanish Stress: Developmental and Theoretical Perspectives. *Language* 64(4). 683-706.

- Hudson Kam, Carla & Elissa Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1. 151-195.
- Jusczyk, Peter & Richard Aslin. 1995. Infants detection of the sound patterns of words in fluent speech, *Cognitive Psychology* 29. 1–23.
- Jusczyk, Peter, Anne Cutler, & Nancy Redanz. 1993. Infants' preference for the predominant stress patterns of English words. *Child Development* 64. 675-687.
- Jusczyk, Peter, Mara Goodman, & Angela Baumann. 1999. Nine-month-olds' attention to sound similarities in syllables, *Journal of Memory & Language* 40. 62–82.
- Jusczyk, Peter, Derek Houston, & Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology* 39. 159–207.
- Kehoe, Margaret. 1997. Stress error patterns in English-speaking children's word productions. *Clinical Linguistics and Phonetics* 11(5). 389-409.
- Kehoe, Margaret. 1998. Support for metrical stress theory in stress acquisition. *Clinical Linguistics & Phonetics* 12(1). 1-23.
- Lightfoot, David. 1999. *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell.
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Niyogi, Partha & Robert Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61. 161-193.
- Nouveau, Dominique. 1994. *Language Acquisition, Metrical Theory, and Optimality: A Study of Dutch Word Stress*. Utrecht: Utrecht University dissertation.

- Pearl, Lisa. 2007. Necessary Bias in Natural Language Learning. College Park: Maryland: University of Maryland dissertation.
- Pearl, Lisa. 2008. Putting the Emphasis on Unambiguous: The Feasibility of Data Filtering for Learning English Metrical Phonology. *Proceedings of the 32nd Annual Boston Conference on Child Language Development (BUCLD 32)*, 390-401. Somerville, MA: Cascadilla Press.
- Pearl, Lisa. 2009. Acquiring Complex Linguistic Systems From Natural Language Data: What Selective Learning Biases Can Do. Ms. University of California, Irvine.
- Pearl, Lisa & Jeffrey Lidz. 2009. When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development* 5(4). 235-265.
- Pearl, Lisa & Amy Weinberg. 2007. Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling. *Language Learning and Development* 3(1). 43-72.
- Perfors, Amy, Joshua Tenenbaum, & Terry Regier. 2006. Poverty of the Stimulus? A rational approach. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada.
- Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14. 19-100.
- Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.
- Regier, Terry & Susanne Gahl. 2004. Learning the unlearnable: The role of missing evidence. *Cognition* 93. 147-155.
- Sakas, William. 2003. A Word-Order Database for Testing Computational Models of Language Acquisition. *Proceedings of the 41st Annual Meeting of the Association for*

Computational Linguistics.

- Sakas, William and Janet Fodor. 2001. The Structural Triggers Learner. In Stefano Bertolo (ed.), *Language Acquisition and Learnability*, 172-233. Cambridge: Cambridge University Press.
- Sakas, William & Eiji Nishimoto. 2002. Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Ms: City University of New York.
- Tesar, Bruce & Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA: The MIT Press.
- Tomasello, Michael. 2006. Acquiring linguistic constructions. In William Damon, Richard Lerner, Deanna Kuhn & Robert Siegler (eds.), *Handbook of Child Psychology*, 255-298. New York: Wiley.
- Turk, Alice, Peter Jusczyk, & LouAnn Gerken. 1995. Do English-learning Infants Use Syllable Weight to Determine Stress? *Language and Speech* 38(2).143-158.
- Vallabha, Gautam, James McClelland, Ferran Pons, Janet Werker, & Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the U.S.* 104(33). 13273-13278.
- Wilson, Michael. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers* 20(1). 6-11.
- Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.
- Yang, C. 2005. On productivity. *Yearbook of Language Variation* 5. 333-370.

A1. Appendix: Child-directed and Adult-directed speech data

The child-directed speech data comprising the input of the constrained learners were taken from the Brent (Brent & Siskind 2001) and Bernstein (Bernstein Ratner 1984) corpora in CHILDES (MacWhinney 2000). The token and type distributions of this corpus are shown below in Table A1. For each n-syllable word class, the frequency of each stress pattern is shown along with at least one example word that has that stress pattern and appeared in the child-directed speech data. Stressed syllables are represented as 1, while unstressed syllables are represented as 0, e.g., the pattern ‘01’ represents an unstressed syllable followed by a stressed syllable, such as in the word *giraffe*. Stress patterns absent from the table have a token and type frequency of 0.

Total: (540505 tokens / 8093 types)		
Words with the same number of syllables: (tokens / types)		
1-syl: (449312 / 4474)	2-syl: (85268 / 2898)	3-syl: (4749 / 476)
Stress Pattern Frequency	Stress Pattern Frequency	Stress Pattern Frequency
1: (373838 / 4420) <i>book, know</i>	11: (11213 / 401) <i>something, snowman</i>	110: (572 / 109) <i>dishwasher, sunglasses</i>
0: (75474 / 54) <i>as, is</i>	10: (66568 / 2236) <i>over, sleeping</i>	101: (3049 / 272) <i>basketball, understand</i>
	01: (7487 / 261) <i>around, himself</i>	100: (689 / 60) <i>wonderful, earlier</i>

		011: (6 / 5) <i>repairman, peroxide</i>
		010: (433 / 30) <i>important, adventure</i>
4-syl: (1008 / 214)		5-syl: (163 / 26)
Stress Pattern Frequency		Stress Pattern Frequency
1101: (1 / 1) <i>identify</i>	0110: (2 / 1) <i>dehydrated</i>	11010: (1 / 1) <i>rhinoceroses</i>
1100: (20 / 8) <i>tasmanian</i>	0101: (18 / 14) <i>congratulate</i>	01100: (1 / 1) <i>electronically</i>
1010: (910 / 161) <i>conversation</i>	10101: (54 / 1) <i>cockadoodledoo</i>	101010: (67 / 8) <i>fantabalocious</i>
1001: (7 / 3) <i>pedialyte</i>	0100: (50 / 26) <i>unfortunate</i>	10100: (39 / 15) <i>personality</i>
		01000: (2 / 1) <i>cooperative</i>
6-syl: (4 / 4)		7-syl: (1 / 1)
Stress Pattern Frequency		Stress Pattern Frequency
100100: (2 / 2) <i>czechoslovakia's</i>	010100: (1 / 1) <i>impossibility</i>	1010100: (1 / 1) <i>unidentifiable</i>
100010: (1 / 1) <i>hypoallergenic</i>		

Table A1. Child-directed speech data.

The adult-directed conversational speech data examined come from the North American English CALLFRIEND corpus (Canavan & Zipperlen 1996). The token and type distributions of this corpus are shown below in Table A2. For each n-syllable word class, the frequency of each stress pattern is shown along with at least one example word that has that stress pattern and appeared in the adult-directed speech data. Stressed syllables are represented as 1, while unstressed syllables are represented as 0, e.g., the pattern ‘01’ represents an unstressed syllable followed by a stressed syllable, such as in the word *giraffe*. Stress patterns absent from the table have a token and type frequency of 0.

Total: (82487 tokens / 4719 types)		
Words with the same number of syllables: (tokens / types)		
1-syl: (68252 / 1868)	2-syl: (11059 / 1780)	3-syl: (2554 / 755)
Stress Pattern Frequency	Stress Pattern Frequency	Stress Pattern Frequency
1: (40249 / 1794) <i>down, fast</i>	11: (1409 / 224) <i>something, sixteen</i>	111: (1 / 1) <i>giordano</i>
0: (28003 / 74) <i>as, is</i>	10: (8445 / 1298) <i>very, smiling</i>	110: (315 / 78) <i>thanksgiving, however</i>
	01: (1205 / 258) <i>around, supposed</i>	101: (553 / 141) <i>saturday, understand</i>
		100: (948 / 270) <i>video, totally</i>
		011: (11 / 6)

		<i>tornado, chicago</i>	
		010: (726 / 259)	
		<i>semester, tomorrow</i>	
4-syl: (468 / 242)		5-syl: (141 / 64)	
Stress Pattern Frequency		Stress Pattern Frequency	
1110: (1 / 1)	1101: (2 / 1)	11100: (2 / 1)	11010: (1 / 1)
<i>relaxation</i>	<i>identified</i>	<i>biometrical</i>	<i>biomechanics</i>
1100: (19 / 13)	1011: (3 / 1)	10100: (90 / 32)	10010: (13 / 8)
<i>priorities</i>	<i>ibuprofen</i>	<i>understandable</i>	<i>radioactive</i>
1010: (235 / 114)	1001: (4 / 3)	01010: (20 / 13)	01000: (15 / 9)
<i>situation</i>	<i>videotape</i>	<i>discrimination</i>	<i>immediately</i>
1000: (25 / 14)	0110: (3 / 1)		
<i>obviously</i>	<i>distributed</i>		
0101: (22 / 7)	0100: (154 / 87)		
<i>relationship</i>	<i>perfectionist</i>		
6-syl: (13 / 10)			
Stress Pattern Frequency			
110100: (1 / 1)	101100: (3 / 2)	101010: (2 / 1)	
<i>postbaccalaureate</i>	<i>heterosexual</i>	<i>bioengineering</i>	
101000: (2 / 2)	100100: (1 / 1)	100010: (1 / 1)	
<i>indistinguishable</i>	<i>spirituality</i>	<i>identification</i>	
010100: (3 / 2)			
<i>suggestopedia</i>			

Table A2. Adult-directed conversational speech data.

Appendix A2. Exceptions to the English grammar in child-directed speech.

The English grammar (following Dresher (1999), who draws from Halle & Vergnaud (1987)) is quantity sensitive (QS) with closed syllables viewed as Heavy (QS-VC-H), has the rightmost syllable as extrametrical (Em-Some, Em-Right), constructs metrical feet starting from the right edge of the word (Ft-Dir-Rt), specifies metrical feet as two syllables in size (B, B-2, B-Syl), and stresses the leftmost syllable in a foot (Ft-Hd-Left). Given child-directed speech data from the Bernstein Ratner (Bernstein Ratner 1984) and Brent (Brent Siskind 2001) corpora from the CHILDES database (MacWhinney 2000), we found that this grammar was only compatible with 73.0% of the data tokens and 62.1% of the data types.

A common set of exceptions involves words that stress the rightmost syllable of the word, thereby violating Em-Right (since extrametrical syllables are not stressed): 81.8% of the token exceptions and 57.1% of the type exceptions involve a word of this kind, and include commonly appearing words such as *herself, himself, instead, backyard, bathtub, football, someone, something, somewhere, sweatshirt, washcloth, birthday, goodbye, onto, alright, balloon, excuse, inside, airplane, almost, always, bathroom, bedroom, cupcake, sometimes, sunshine, haystack, myself, peanut, playpen, snowman, okay, eighteen, playhouse, seesaw, baseball, downstairs, goldfish, mailbox, outfit, reindeer, sweetheart, toothbrush, rainbow, oatmeal, outside, above, across, again, ahead, because, before, enough, giraffe, return, away, alone, around, supposed, tonight, dinosaur, pokemon, microphone, microwave, and kangaroo.*

Another common set of exceptions are words that contain an unstressed internal syllable that would be viewed as heavy (closed (VC), long (VV) or superlong (VVC)). Because this syllable could not be unstressed due to extrametricality since it is not at a word edge, such words present a problem for the quantity sensitive (QS) setting that views closed syllables as heavy (QS-VC-H); specifically, these internal syllables should be prominent due to their syllable weight, but are nonetheless unstressed. These words accounted for 11.4% of the token exceptions and 22.2% of the type exceptions, including commonly appearing words such as *wonderful*, *certainly*, *indians*, *oranges*, *caterpillar*, *watermelon*, and *everybody*. There were, of course, often-used words that had this issue as well as having stress on the rightmost syllable, such as *basketball*, *understand*, *yesterday*, *underneath*, *neighborhood*, *overalls*, *radio*, *studio*, *applesauce*, *somersault*, *waterfall*, *butterfly*, *seventeen*, *anymore*, *everyone*, *everything*, *everywhere*, *lollipop*, *pussycat*, *anyway*, and *pattycake*.

Notably, many of these exceptions have the form of compound words. It could be very useful to children to realize that compound words are likely to be exceptions when they are trying to identify the adult grammar for English.

Appendix A3. Selective learning biases: Unambiguous data

Pearl (2008, 2009) found that probabilistic learners biased to learn only from data perceived as unambiguous were successful at acquiring the English grammar for the parametric system described here. One difficulty with this strategy is that children must somehow identify which data are unambiguous in the input. A data point is defined as unambiguous with respect to a given parameter value (e.g., an unambiguous data point

for Em-None may be ambiguous for the foot directionality parameter). Two proposals for how children could identify unambiguous data are that (1) they look for certain configurations in the input, called *cues* (Dresher 1999, Lightfoot 1999), and (2) they *parse* a data point to determine if only a single parameter value for a given parameter yields a stress contour that matches the observed stress contour (Fodor 1998b, Sakas & Fodor 2001).

Cues for each value of the metrical phonology system are given in Table A3, with an example of each cue in parentheses after the description of the cue. Note that most cues depend on the current state of the child’s knowledge (e.g., see the cues for QS, Ft-Dir-Left, B-Syl, and Ft-Hd-Left). Note that these cues are not the cues proposed in Dresher (1999), but are designed in the same spirit – to identify highly informative data. Unlike some of the cues in Dresher’s proposal, all these cues can be identified within a single data point. This is in contrast to cues that operate over multiple data points. Not needing to compare multiple data points may be desirable if the child is simply extracting information from the current data point and integrating that information into her knowledge of the parametric system, rather than explicitly comparing the current data point to items already in the lexicon. In addition, cues are proposed not just for those parameter values that could be viewed as marked, but also for parameter values that could be viewed as the default option.

Parameter	Cue
Value	
QI	Unstressed internal VV(C) syllable (...VV...)

QS	<i>Em-None or Em unknown:</i> 2 syllable word with 2 stresses (<u>VV</u> <u>VC</u>) <i>Em-Some:</i> 3 syllable word, with 2 adjacent syllables stressed (<u>VC</u> <u>VV</u> <u>VC</u>)
QS-VC-L	Unstressed internal VC syllable (...VC...)
QS-VC-H	<i>Em-None or Em unknown:</i> 2 syllable word with 2 stresses, one or more are VC syllables (<u>VV</u> <u>VC</u>) <i>Em-Some:</i> 3 syllable word, with 2 adjacent syllables stressed, one or more are VC syllables (<u>VC</u> <u>VV</u> <u>VC</u>)
Em-None	Both edge syllables are stressed (<u>V</u> ... <u>VC</u>)
Em-Some	Union of Em-Left and Em-Right cues
Em-Left	Leftmost syllable is Heavy and unstressed (H...)
Em-Right	Rightmost syllable is Heavy and unstressed (...H)
Ft-Dir-Left	<i>QI or Q-unknown, Em-None/Left or Em unknown:</i> 2 stressed adjacent syllables at right edge (... <u>VC</u> <u>V</u>) <i>QI or Q-unknown, Em-Right:</i> 2 stressed adjacent syllables followed by unstressed syllable at right edge (... <u>VC</u> <u>V</u> <u>VV</u>) <i>QS, Em-None/Left or Em unknown:</i> stressed H syllable followed by stressed L syllable at right edge (... <u>H</u> <u>L</u>) <i>QS, Em-Right:</i> stressed H syllable followed by stressed L syllable followed by unstressed syllable at right edge (... <u>H</u> <u>L</u> H)
Ft-Dir-Rt	<i>QI or Q-unknown, Em-None/Right or Em unknown:</i> 2 stressed adjacent syllables at left edge (<u>VC</u> <u>V</u> ...) <i>QI or Q-unknown, Em-Left:</i> unstressed syllable followed by 2 stressed

	adjacent syllables at left edge (VC <u>V</u> <u>VV</u> ...)
	<i>QS, Em-None/Right or Em unknown</i> : stressed L syllable followed by stressed H syllable at left edge (<u>L</u> <u>H</u> ...)
	<i>QS, Em-Left</i> : unstressed syllable followed by stressed L syllable followed by stressed H at left edge (H <u>L</u> <u>H</u> ...)
Unb	<i>QI or Q-unknown</i> : 3+ unstressed syllables in a row (...VC VV VC...) <i>QS</i> : 3+ unstressed Light syllables in a row (...L L L)
B	Union of B-2 and B-3 cues
B-2	<i>QI or Q-unknown</i> : 3+ syllables in a row, every other one stressed (... <u>VC</u> VV <u>VC</u> ...) <i>QS</i> : 3+ Light syllables in a row, every other one stressed (... <u>L</u> L <u>L</u> ...)
B-3	<i>QI or Q-unknown</i> : 4+ syllables in a row, every third one stressed (... <u>V</u> VC VV <u>V</u> ...) <i>QS</i> : 4+ Light syllables in a row, every third one stressed (... <u>L</u> L L <u>L</u> ...)
B-Syl	<i>QI or Q-unknown</i> : Union of QI B-2 and QI B-3 cues <i>QS, B-2</i> : 2 adjacent syllables, one stressed Heavy and one unstressed Light (... <u>H</u> L...) <i>QS, B-3</i> : 3 adjacent syllables, 2 unstressed Light preceding a stressed Heavy or following a stressed Heavy (... <u>H</u> L L...), (...L L <u>H</u> ...)
B-Mor	<i>Em-None or Em-unknown</i> : 2 syllable word with both syllables Heavy and stressed (<u>H</u> <u>H</u>) <i>Em-Some</i> : 3 syllable word with 2 adjacent syllables Heavy and stressed (L <u>H</u> <u>H</u>)

Ft-Hd-Left	<i>Em-None or Em-unknown</i> : Leftmost syllable is stressed (<u>VC</u> ...)
	<i>Em-Left</i> : 2 nd from leftmost syllable is stressed (VV <u>VC</u> ...)
Ft-Hd-Rt	<i>Em-None of Em-unknown</i> : Rightmost syllable is stressed (... <u>VC</u>)
	<i>Em-Right</i> : 2 nd from rightmost syllable is stressed (... <u>VC</u> VV)

Table A3. Cues for metrical phonology parameter values. Some cues may depend on the child's current knowledge state, represented in *italics*. For example, the cue for QS depends on what is known about extrametricality (*Em-None/Em-Some/Em-unknown*).

The parsing method involves the child using the structure-assigning ability of parsing that is presumed to be used already during language comprehension (Fodor 1998a, 1998b, Sakas & Fodor 2001). The parsing instantiation examined in Pearl (2008, 2009) tries to analyze a data point with “all possible parameter value combinations”, conducting an exhaustive search of “all parametric possibilities” (Fodor 1998b). For this kind of parsing, a successful parameter value combination will generate a stress contour that matches the observed stress contour of the data point - this is then a successful parse of the data point. For instance, the combination QI, Em-None, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left is able to generate the stress contour [stressed unstressed stressed] for the word *afternoon*. Since the stress contour the child would encounter for *afternoon* matches this stress contour (*afternoon*), this combination can successfully parse this data point.

If all successful parses use only one of the available parameter values for a given parameter (e.g. Em-None of the extrametricality values), that data point is viewed as unambiguous for that parameter value. Data points that can be parsed with multiple

parameter values of the same parameter (e.g., Ft-Hd-Left and Ft-Hd-Rt for the foot headedness parameter) are considered ambiguous. These ambiguous data points are filtered out of the child's intake for that parameter value (e.g., foot headedness) by the child's unambiguous data learning bias.

As an example of this parsing method in action, suppose the child encounters *afternoon*, and successfully recognizes two pieces of information: (1) the syllables are *af* (VC), *ter* (VC), and *noon* (VVC), and (2) the associated stress contour is VC VC VVC. A parsing child would try to generate the observed stress contour with all available parameter value combinations and come up with five that are successful:

- (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-VC-L, **Em-None**, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left
- (d) QS, QS-VC-L, **Em-None**, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Rt
- (e) QS, QS-VC-L, **Em-None**, Ft-Dir-Left, Unb, Ft-Hd-Left)

All these successful parses share Em-None, meaning that Em-None was required for a successful parse. The child then perceives this data point as unambiguous for Em-None.

It turns out that a general class of probabilistic learners using either of these identification methods is guaranteed to succeed at selecting the parameter values that comprise the English grammar (i.e., the adult English grammar described in the main text) when trained on the same corpus of English child-directed speech used for the

unbiased probabilistic learners examined here (see Pearl (2008, 2009) for details).

However, these biased learners require that certain parameter-setting order constraints be obeyed. For cues, these are as follows: (a) QS-VC-H before Em-Right, (b) Em-Right before B-Syl, and (c) B-2 before B-Syl. That is, (a) the child must determine that VC syllables are treated as Heavy (QS-VC-H) before determining that the rightmost syllable is extrametrical (Em-Right), (b) the child must determine that the rightmost syllable is extrametrical (Em-Right) before determining that a metrical foot's size is determined by the number of syllables it contains (B-Syl), and (c) the child must determine that metrical feet are two units in size (B-2) before determining that a metrical foot's size is determined by the number of syllables it contains (B-Syl).

For parsing, there are three groups such that the first one must be set before the second one, and the second one must be set before the third one:

Group 1: QS, B, Ft-Hd-Left

Group 2: Ft-Dir-Rt, QS-VC-H

Group 3: Em-Some, Em-Right, B-2, B-Syl

So, first the parsing child must determine that the language is quantity sensitive (QS), that metrical feet are of some arbitrary bounded size (B), and that metrical feet are headed on the left (Ft-Hd-Left) before determining any of the other parameters of the English grammar. Then, the parsing child must determine that metrical feet are constructed starting from the right edge of the word (Ft-Dir-Rt) and VC syllables are treated as Heavy (QS-VC-H). Finally, the parsing child can determine that the rightmost syllable is

extrametrical (Em-Some, Em-Right) and metrical feet are two syllables in size (B-2, B-Syl).

If the child does not follow these parameter-setting order constraints, whether that child is using cues or parsing to identify unambiguous data, the child will not observe distributions of unambiguous data that favor the adult English grammar's parameter values. Thus, to succeed, a child trying to acquire the adult English grammar in this parametric system must not only know to learn from unambiguous data alone, but must also either have or derive the knowledge of these parameter-setting order constraints. See Pearl (2007, 2009) for discussion of how some or all of these constraints may be derived from properties such as data saliency, data quantity, and default values for the parametric system, depending on the identification method.