

**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 Pearl

Department of Cognitive Sciences

University of California, Irvine

[lpearl@uci.edu](mailto:lpearl@uci.edu)

## **Abstract**

Parametric systems have been proposed as models of knowledge representations about language, often as a way to explain children's rapid acquisition of linguistic knowledge. Given this aim, it seems reasonable to examine if children with knowledge of parameters could in fact acquire the adult system from the data available to them. Recent computational modeling work has suggested that children could acquire a parametric implementation of the English metrical phonology system if selective learning biases are coupled with probabilistic learning. The present modeling study supports the necessity of a learning bias by showing that unbiased probabilistic learners will fail on this same case study, due to the nature of the data available to English children. Special attention is given to the model's 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.

## 1. Introduction: Acquisition of systematic knowledge

Knowledge of language includes many kinds of systematic knowledge, such as phonology, morphology, and syntax. Successful acquisition involves discovering this systematic knowledge for the native language. Theoretical research traditionally provides a description of the knowledge to be acquired, identifying what children must know. Experimental research often provides the time course of acquisition, identifying when children attain this knowledge. To understand how acquisition works, we can create computational models that draw on the linguistic representations from theoretical research and the trajectory of acquisition from experimental research. With this information, computational models can be grounded both theoretically and empirically, and thus modeling can concretely examine questions about the acquisition mechanism. Specifically, within a model, we can precisely manipulate some part of the acquisition mechanism and see the results on 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 the linguistic system of metrical phonology, using child-directed speech data as input, and keeping in mind the 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 model are more difficult to do with children – such as controlling the hypotheses they entertain, 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.

### 1.1. Linguistic systems

Acquisition is not so easy if we believe children are acquiring a complex system (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, 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)); second, it is used to explain how children converge quickly on the complex linguistic knowledge they seem to attain. 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 (e.g. metrical phonology knowledge: see Hochberg 1988). To build the correct system as rapidly as children do, it is then 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. So, under this view, the basic purpose of children having prior knowledge of linguistic parameters is to make the acquisition of a complex system possible in the time frame children have to do it. Given this, it seems reasonable to ask if a given proposal for prior knowledge makes the complex system it is designed to help acquire actually acquirable.

## 1.2. Parametric metrical phonology

One proposal of children's prior knowledge is a parametric system, and one domain for which it has been proposed is metrical phonology (Halle & Vergnaud 1987, Hayes 1995). To acquire a parametric system, children must view the encountered data as the output of that 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 *elephant* (stressed syllables will be indicated by underlining henceforth), which has the stress contour [stressed unstressed unstressed]. 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. *e*, *l*, *t*), (b) if the individual syllables matter (e.g. *el*, *phant*), (c) if rhyming matters (e.g. *el* does not rhyme with *phant*), (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.

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 not really solved. 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 a parametric system of metrical phonology for English from the input available to English children. Prior computational modeling work (Pearl 2008, submitted) has suggested that the adult parametric system is acquirable from English child-directed speech as long as the child employs a selective learning bias while also being sensitive to the probability distributions in the data. Since this system is acquirable from realistic data, we have a mark in favor of this proposal of knowledge representation. However, an open question is whether a selective learning bias is necessary for a child who can leverage the information available in probability distributions. That is, do children need an additional learning bias to succeed?

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. 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 Hayes (1995) and the parameterization used in Drescher (1999) that draws from Halle & Vergnaud (1987). 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). There are five main binary parameters and four binary sub-parameters, yielding 156 grammars in the hypothesis space. The resultant grammars concern only whether syllables are stressed or unstressed, and not how much stress syllables receive compared to other syllables. Moreover, these grammars do not describe interactions with the morphology system, due to considerations of the child's likely initial knowledge state when acquiring the metrical phonology system. Experimental research (Jusczyk, Cutler, & Redanz 1993, Turk, Jusczyk, & Gerken 1995) suggests that children under a year old may already have some knowledge of aspects of metrical phonology. Kehoe (1998) suggests that children already know several parameter values of the English system by 22 months. It seems unlikely that children at these ages have extensive knowledge of their language's morphology, and so they may not yet hypothesize interactions between the morphology system and the metrical phonology system. We thus proceed with the following assumption: 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 system. 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 model here uses 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 unbiased probabilistic learning models that attempt to acquire the English grammar from data distributions estimated from English child-directed speech. Perhaps surprisingly, they fail to acquire the correct grammar with any reliability. Following that, we identify the source of their failure and find it is intrinsic to the English data set. For this reason, *no* unbiased probabilistic model would succeed at this task. We conclude with discussion of implications for the acquisition mechanism.

## 2. The parameters of the system

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

### (2) metrical phonology analysis for *elephant*

(H	L)	<H>
<i>el</i>	<i>e</i>	<i>phant</i>

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 (adapted from Dresher (1999) and Hayes (1995)). For a detailed description of each of the parameters and their interactions with each other, see Pearl (2007).

One parameter, quantity sensitivity, concerns whether all syllables are identical, or differentiated by syllable rime weight. 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)eavy and (L)ight syllables. Long vowel syllables (VV) 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 *company*.

(3) QS and QI analyses of <i>company</i>								
QS analysis	L/H	L	H	QI analysis	S	S	S	
syllable rime	VC	V	VV					
syllable structure	CVC	CV	CCVV					
syllables	<u>com</u>	pa	ny		<u>com</u>	pa	ny	

Syllables classified as Heavy should receive stress, but sometimes do not 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 *company*, while (4b) shows an Em-None analysis for *afternoon*.

(4a) Em-Some analyses

Em-Left				Em-Right			
syllable class	<L>	(H)		(H	L)	<H>	
syllable rime	V	VC		VC	V	VV	
syllables	<i>gi</i>	<i>raffe</i>		<u>com</u>	pa	ny	

(4b) An Em-None analysis

syllable class	(L	L)	(H)
syllable rime	VC	VC	VV
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 feet 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 feet construction from the left (Ft-Dir-Left): (L L H  
(5b) Start metrical feet 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, then the metrical foot encompasses all the non-extrametrical syllables in the word. Some example Unb analyses are shown in (6).

(6) Unb analyses

- (a) Differentiated syllables, building feet from the left (Ft-Dir-Left)  
(L L L) (H L)  
(b) Differentiated syllables, building feet from the right (Ft-Dir-Rt)  
(L L L H) (L)

(c) (Un)differentiated syllables, building feet from either direction

(L L L L L)  
(S S S S S)

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. Example (7) demonstrates different bounded analyses, with various combinations of these parameter values.

(7) Bounded analyses of four syllable sequences

B-2, B-Syl

(a) (L H) (L L)  
(b) (H H) (L L)  
(c) (S S) (S S)

B-2, B-Mor

(d) mora analysis                    μ μ μ μ μ μ  
syllable classification        (H) (H) (L L)

Once the metrical feet are formed, the feet headedness parameter determines which syllable within a foot is stressed. Feet headed on the left have the leftmost syllable of the foot stressed (Ft-Hd-Left) while feet headed on the right have the rightmost syllable of the foot stressed (Ft-Hd-Rt). Example (8) shows samples of both analyses.

(8) Ft-Hd-Left and Ft-Hd-Rt analyses for (H L) (L)

(a) Ft-Hd-Left: (H L) (L)  
(b) Ft-Hd-Rt: (H L) (L)

These five parameters (quantity sensitivity, extrametricality, feet directionality, boundedness, feet headedness) and their sub-parameters (VC-H/L, Em-Left/Right, B-2/3, B-Syl/Mor) yield 156 grammars in the hypothesis space. 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.

(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.

(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	<u>cu</u>	<u>cum</u>	ber

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

syllable class	(H)	(H)	<H>
syllables	<u>cu</u>	<u>cum</u>	ber

### 3. English

The particular language considered in this modeling study is English, which has the following parameter values: 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 that favor the incorrect parameter values for English, again making acquisition non-trivial. More specifically, the English grammar is incompatible with approximately 27% of the available data by tokens (each instance of a word is counted once), and with approximately 37% by types (each word is counted once no matter how many times it occurs). That is, for 27% of the data tokens (and 37% of the types), the child can only conclude that parameter values other than the English values are responsible for generating the data point. So, these data points are noise with respect to the English grammar. A reasonable question is if a grammar incompatible with such a large portion of the data is really the right grammar. While there obviously must be some way to deal with these exceptional data, a grammar that can reliably cover a majority of the data is still a useful grammar for children to have. 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. Since many of the exceptional data are due to interaction with the morphological system, no grammar in the hypothesis space (which does not contain interactions with morphology) will be able to cover much more than this in the data.

A third reason is that previous computational modeling research (Pearl 2008, submitted) has found that the English grammar can be acquired from child-directed English speech data if the child has a bias to learn only from unambiguous data 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 models 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 (2008) for details), in addition to restrictions on the order in which parameter values could be set. However, if unbiased models are unsuccessful, we can examine why they fail and whether the problem that afflicts these models 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.

#### 4. The model

##### 4.1. The model's input

The model'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 six months and two years old. This age range was estimated as the time period when children might be setting the parameters of the metrical phonology system under consideration, given that several parameters of this system seem to be known by 22 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;6 and 1;0. In total, this yielded 540505 words of orthographically transcribed child-directed speech, consisting of 8093 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 and the associated stress contour was estimated by referencing the CALLHOME American English Lexicon (Canavan *et al.* 1997) and the MRC Psycholinguistic Database (Wilson 1988). In cases of conflict, the CALLHOME database was given preference. Words not present in these two databases of pronunciation were given a pronunciation consistent with the conventions in the CALLHOME database – such words were usually child-register words, e.g. *booboo*. See the Appendix for a detailed summary of the corpus.

The model 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 period, based on the estimates for a three year period 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 forthcoming). 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 models presented use incremental/online update procedures, meaning that they extract information from the data as the data come in. This is in contrast to models that must store all the data encountered to analyze together at some future point (e.g. Perfors, Tenenbaum & Regier 2006; Goldwater, Griffiths, & Johnson 2007; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum 2007, forthcoming). Often such models 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 forthcoming), and these latter studies are more likely to use algorithms that 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 model 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 structure and the stress contour. For example, the word *cucumber* has the syllable structure ‘VV VC VC’ (based on syllable rime) 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 model 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 model 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 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 model 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 model 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	<u>cu</u>	<u>cum</u>	<u>ber</u>

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	<u>cu</u>	<u>cum</u>	<u>ber</u>

mismatch: punish all

When the probability for one parameter value approaches 0.0 or 1.0, the model 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 model 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). Ideally, after a reasonable number of English data points, the model will set the correct values for the English grammar.

The unbiased models considered here vary with respect to how they implement the reward/punishment component of the update procedure. One model 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  $\gamma$  that determines how liberal the model is. The larger  $\gamma$  is, the more probability the model shifts for a single data point.

(12) Linear Reward-Penalty Scheme

$p_v$  = previous probability of parameter value

$p_o$  = previous probability of opposing parameter value

(a) generated stress contour matches observed stress contour (reward)<sup>1</sup>

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

$$p_{Onew} = 1 - p_{vnew}$$

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

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

$$p_{Onew} = 1 - p_{vnew}$$

---

<sup>1</sup> 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.

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 and that grammar fails to generate the observed stress contour. The QI value (and all other participating values) are punished. Suppose  $\gamma$  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 model 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 model 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,  $\alpha$  and  $\beta$  (see (13)). Setting both of these values to 0.5 initially biases the model to favor neither parameter value, and also to prefer probabilities closer to the endpoints (0.0 and 1.0) over probabilities in the middle (e.g. 0.5). This means the learner has no initial preference for a parameter's value, but is initially biased to choose one over the other as data come in. 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

$p_o$  = previous probability of opposing parameter value

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

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

$$p_{Onew, \text{normalized}} = \frac{p_o}{p_{vnew} + p_o}$$

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.

Ideally, either one or both of these models will be able to succeed at acquiring the correct English grammar from English child-directed speech. However, neither model 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 model to analyze larger quantities of data simultaneously. Instead, the model 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 model 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 model 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 model 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 model 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 model. If the NPLearner uses a  $\gamma$  of 0.01, the new probability of QI is 0.495 and the new probability of QS is 0.505. If the BayesLearner model 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 model 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, each model includes the assumption that children can successfully identify words in fluent speech. This does not seem unreasonable as word segmentation research by Jusczyk and colleagues suggests that children as young as seven months can identify words in fluent speech successfully (Jusczyk & Aslin 1995, Jusczyk, Houston, & Newsome 1999). In addition, each data point was pre-divided into syllables, with individual syllables identified by rime as type VV, VC, or V. Thus, the models 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). Thirdly, the models did not call the update procedure if a monosyllabic word was encountered, as monosyllabic words do not have a stress contour per se. Instead, monosyllabic words were effectively ignored.<sup>2</sup> Fourthly, the models 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.

#### 4.5. Model parameters and simulations

The four models – NPLearner, BayesLearner, Count NPLearner, and Count BayesLearner – were run on the input set, which was generated from the English child-directed speech distributions. 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 model 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 6 months (1,666,667).

### 5. Results and discussion

Table 1 shows the percentage of the trials each model converged on the English grammar. When multiple parameter values are used within a model (e.g.  $c = 2, 5, 7, 10, 15, \text{ or } 20$  for the counting variants), the average percent convergence is given. The most striking aspect of these results is the extreme rarity with which these unbiased models

---

<sup>2</sup> Notably, simulations with models that processed monosyllabic words never converged on the English grammar due to the extrametricality parameter. Specifically, 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. Therefore, all the simulation results reported are from models that ignore monosyllabic words for the purposes of acquiring the metrical phonology system.

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.

Unbiased Model	% English Convergence
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 modeling results.

How do we interpret this lack of convergence for unbiased probabilistic models? Recall that the biased model from Pearl (2008, submitted) guaranteed convergence so long as the child learned only from unambiguous data and set the parameters in a particular order. Thus, convergence in a small fraction of the trials here certainly does not look like the desired acquisition behavior. If we look closer at the modeling results obtained here, we can see what kind of errors the unbiased models are making. It seems in general that these models 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.

(14) Examples of incorrect grammars selected by unbiased models

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

Obviously, this exceptionally poor performance was not the behavior we were looking for from these acquisition models. The question then is whether the problem is with these models in particular, or if there is some underlying issue that will cause all unbiased probabilistic models to fail. If the problem is with these models, then we simply need to try better models. However, if the problem is inherent to the acquisition scenario somehow, then no unbiased probabilistic model can be successful.

Let us examine the acquisition scenario in more detail. The hypothesis space contains 156 grammars: the English grammar 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 does it fare compare to the other 155? 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 (disregarding the frequency with which words

appear in the input), English is less compatible than 56 other grammars. The point is simple, and seems rather surprising: English is not the optimal grammar for the English data set.

We can then ask if the unbiased models are converging on the grammars that are in fact more optimal for the data set they are given. English is compatible with 72.97% of the tokens in the English child-directed data, and with 62.14% of the types. The average compatibility of the grammars these unbiased models select is higher: 73.56% by data tokens and 63.30% by data types. The unbiased models are therefore doing what they should: identifying the more optimal grammars in the hypothesis space, given the input data. In short, the failure of these unbiased models is not because they cannot find the more optimal grammars given the English input; the failure is *because* they find the more optimal grammars given the English input.

Because the English grammar is non-optimal for the English input (and in fact, barely in the top third), it is less surprising that unbiased probabilistic models do not converge on the English grammar reliably. Unbiased models are geared to find the more optimal hypotheses in a hypothesis space. Therefore, they do not converge on the English grammar.

However, this means the behavior of unbiased models does not accord with the behavior we expect to see in English children, i.e. converging on the English grammar. One way to solve this is to believe that English children are not unbiased probabilistic learners. Instead, to acquire this parametric system, children must incorporate some biases into their acquisition process. In particular, English children must have some bias that makes English a far more optimal grammar for the data they encounter.

## 5.2. Bias on the hypothesis space

One bias English children may have is on their hypothesis space of possible grammars. Some of the parameters considered in the metrical phonology system here are related to rhythmic properties of English that children may have already acquired by the time they are acquiring the metrical phonology system. Experimental evidence from Jusczyk, Cutler, & Redanz (1993) suggests that English infants prefer strong-weak bisyllables (e.g. *baba*) over weak-strong bisyllables (e.g. *baba*). This may bias the English child to favor metrical feet headed on the left (Ft-Hd-Left) over metrical feet headed on the right (Ft-Hd-Rt), which is the correct preference for English. Experimental evidence from Turk, Jusczyk, & Gerken (1995) and experiments described in Gerken & Aslin (2005) suggest that English infants are sensitive to syllable structure when determining stress. This may bias the English child to favor quantity sensitive (QS) over quantity insensitive (QI), which is the correct preference for English. If we take the strongest starting point and assume English infants may already be aware that English is quantity sensitive with metrical feet headed on the left (QS, Ft-Hd-Left), the hypothesis space of possible grammars is considerably smaller. Instead of 156 grammars, English learners with this prior knowledge would need to select from only 60 grammars.

We can see if this bias on the hypothesis space is enough to yield more reliable convergence on the English grammar using the probabilistic models from before, NP\_Learner, BayesLearner, Count NP\_Learner, and Count BayesLearner. To simulate the learner's prior knowledge, the probabilities of QS and Ft-Hd-Left are initially 1.0 (rather than 0.5). This means that these values are always chosen to generate a grammar during

data processing. Table 2 shows the percentage of the trials each model converged on the English grammar. When multiple parameter values are used within a model (e.g.  $c = 2, 5, 7, 10, 15,$  or  $20$  for the counting variants), the average percent convergence is given. While there is some improvement over the unbiased models (three of the four manage to converge on the English grammar at least some of the time), the models here still fail to converge on the English grammar with any reliability; the highest percentage of English convergence is achieved by the Count BayesLearner, which still chooses the English system less than 2% of the time.

Model with Hypothesis Space Bias	% English Convergence
NPLearner	
$\gamma = 0.001, 0.0025, 0.01,$ or $0.025$	0.000
BayesLearner	0.100
Count NPLearner	
$\gamma = 0.001, 0.0025, 0.01,$ or $0.025$	
$c = 2, 5, 7, 10, 15,$ or $20$	1.650
Count BayesLearner	
$c = 2, 5, 7, 10, 15,$ or $20$	1.780

Table 2. Modeling results for hypothesis space bias.

Though we shrank the hypothesis space of grammars from 180 to 60 with a bias on the hypothesis space, we saw very little improvement in acquisition. If we look at the compatibility of the other grammars compared to the English grammar, the reason for this failure becomes apparent. English is less compatible than 17 other grammars with respect to both the English data tokens and data types. Again, the English grammar is not the optimal grammar for this data, even within the restricted hypothesis space. As before, it is barely in the top third. This bias on the hypothesis space apparently did not cause the English grammar to be more optimal compared to competing grammars.

### 5.3. Bias on the data intake

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, submitted). In particular, the model learned only from the subset of the available input viewed as unambiguous. A general class of probabilistic models was *guaranteed* to succeed as long as the parameters were acquired in particular orders. Obviously, a guarantee of successful convergence on the English grammar is far superior to the performance of the models we have seen here, both unbiased and those with a bias on the hypothesis space.

The reason why the data intake bias works is because the unambiguous data favor the English parameter values when the parameters are acquired in particular orders. So, if the parameters are acquired in one of those orders, the English grammar is the optimal grammar for the unambiguous data. In that case, a probabilistic learning algorithm that prefers the optimal grammar will converge on the English grammar. Thus, a probabilistically learning child with a bias to learn only from unambiguous data would

succeed, provided that child had knowledge of the appropriate 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 or other learning biases the child has (see Pearl (submitted) for discussion).

Preliminary simulations have suggested that it is not the ordering alone that causes the English grammar to be optimal – when parameters are set in similar orders by the unbiased models described here, there is still no reliable convergence on the English grammar. Moreover, the few times these models 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 is learning only from unambiguous data, and thereby ignoring ambiguous data during acquisition, which is the data intake bias.

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 have the correct distributions to lead the child to the English system, 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. In general, 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.

## 6. Implications for acquisition

Of course, the unambiguous data intake bias discussed above is only one bias that seems to cause English to be the optimal grammar for the given data. There may well be others that accomplish the same thing. The crucial idea is that some kind of bias is needed to produce the acquisition behavior we see in children, an idea noted by several researchers for other case studies in acquisition (e.g. English anaphoric *one*: Regier & Gahl 2004, Pearl & Lidz forthcoming; structure-dependence of syntactic rules: Perfors, Tenenbaum, & Regier 2006). The present study reinforces this for acquiring parametric systems such as the metrical phonology one considered here. The exact nature of the necessary bias can be investigated through computational modeling studies, such as Perfors, Tenenbaum, & Regier (2006) which uses a simplicity bias on the hypothesis space, and Regier & Gahl (2004), Pearl (2008, submitted), and Pearl & Lidz (forthcoming) 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, submitted; Pearl & Lidz forthcoming) or domain-general (Regier & Gahl 2004, Perfors, Tenenbaum, & Regier 2006, Pearl & Lidz forthcoming).

In general, modeling studies can tell us about the mechanism of acquisition, both the process and the prior knowledge required for success. To understand acquisition (as opposed to general learnability), models that learn from realistic data and that consider psychological plausibility in their algorithms are more likely to be informative. Here, the models learned from realistic distributions of English child-directed speech and used online probabilistic algorithms. From this, we discovered that acquiring the correct grammar from realistic data requires some additional bias that makes the English grammar more optimal for the dataset. One bias known to work is to learn only from maximally informative (i.e. unambiguous) data (Pearl 2008, submitted); while the bias

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). However, there may be other biases that would be successful on this case study. For instance, a child may have a bias on the intake to learn only from data that appear to be systematic or productive (Yang 2005).

As another example, perhaps if the hypothesis space is structured differently, e.g. consisting of constraints that must be ranked (Tesar & Smolensky 2000) rather than parameters that must be set, then the English grammar will be more optimal. If it is not, then we have stronger support for the necessity of a bias. If it is, then we have support for one type of hypothesis space representation over another. This is because a hypothesis space that allows a child to converge on the English grammar without a bias is likely to be favored on grounds of simplicity compared to a hypothesis space that requires the child to have a bias. In a similar vein, if we discover that there is no bias that will allow a child to converge on the English grammar in one hypothesis space representation, then that representation is to be disfavored compared to one that allows the child to converge on the correct grammar (whether or not it requires a bias to do so).

We can also examine the robustness of the currently successful acquisition biases, such as the unambiguous data bias, for other acquisition problems. For example, Pearl & Lidz (forthcoming) suggests that learning only from unambiguous data may not be ideal for some acquisition scenarios due to data sparseness. However, some restriction on which ambiguous data are learned from is helpful, i.e. learning only from a particular subset of the ambiguous data in addition to learning from the unambiguous data.

## 7. Conclusion

Acquiring parametric linguistic systems – such as the metrical phonology system considered here – may require more than the ability to leverage the probabilistic information available in the data, given the data children encounter. One way to acquire the correct knowledge is to incorporate biases into the acquisition mechanism. Computational modeling provides a vital tool for examining this possibility, which is difficult to test using traditional experimental techniques. In a model, we control what hypotheses the (simulated) children consider, what data they learn from, and how they alter their beliefs in competing hypotheses. This can allow us to understand how children solve the acquisition problems that they do. The right information may well be in the data – we just have to figure out how children find it.

## Acknowledgements:

To be included.

## 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.
- Bush, Robert & Frederick Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 58. 313-323.
- 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.
- 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 & 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. Forthcoming. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. *Cognitive Science*.
- 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.
- 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. 1998. Support for metrical stress theory in stress acquisition. *Clinical Linguistics & Phonetics* 12(1). 1-23.
- 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.
- 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. Submitted. Acquiring Complex Linguistic Systems From Natural Language Data: What Selective Learning Biases Can Do. Ms. University of California, Irvine.
- Pearl, Lisa & Jeffrey Lidz. Forthcoming. When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*.
- 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 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 41<sup>st</sup> 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 speech data

The child-directed speech data comprising the input 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. 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 1: (373838 / 4420) 0: (75474 / 54)		Stress Pattern Frequency 11: (11213 / 401) 10: (66568 / 2236) 01: (7487 / 261)	
		Stress Pattern Frequency 110: (572 / 109) 101: (3049 / 272) 100: (689 / 60) 011: (6 / 5) 010: (433 / 30)	
4-syl: (1008 / 214)		5-syl: (163 / 26)	
Stress Pattern Frequency 1101: (1 / 1) 1100: (20 / 8) 1010: (910 / 161) 1001: (7 / 3)		Stress Pattern Frequency 11010: (1 / 1) 10101: (54 / 1) 10100: (39 / 15)	
0110: (2 / 1) 0101: (18 / 14) 0100: (50 / 26)		01100: (1 / 1) 01010: (67 / 8) 01000: (2 / 1)	
6-syl: (4 / 4)		7-syl: (1 / 1)	
Stress Pattern Frequency 100100: (2 / 2) 100010: (1 / 1)		Stress Pattern Frequency 1010100: (1 / 1)	
010010: (1 / 1)			

Table A1. Child-directed speech data.