Learning English Metrical Phonology: When Probability Distributions Are Not Enough*

Lisa Pearl
University of California, Irvine

1. Introduction: The Tricky Business of Language Acquisition

Language involves knowledge of multiple complex linguistic systems, such as phonology, morphology, and syntax. For each system, the process of acquisition requires children to discover the underlying components of that system, which native speakers use to generate the observable data. Theoretical research traditionally provides a description of the knowledge to be acquired, identifying what children must know in order to have achieved the same knowledge that adults have. Experimental research often provides the time course of acquisition, identifying when children attain various pieces of this knowledge. Computational modeling research can then draw on the linguistic representations from theoretical research and the trajectory of acquisition from experimental research to examine questions about the process of acquisition. More specifically, within a model, we have complete control over the acquisition mechanism so we can precisely manipulate some part of it and see the results of that manipulation on the acquisition process.

The focus of this paper is the acquisition of the complex linguistic system of metrical phonology, using a model that learns from child-directed speech input and keeps in mind restrictions on the time course of acquisition and children's cognitive limitations. If we believe the model captures the salient aspects of acquisition, the results that come from the manipulations of this model inform us about the nature of acquisition in children. From a practical standpoint, some manipulations that are easy to do in a model are difficult to do with children, e.g. controlling the hypotheses children entertain, the data they learn from, and the way they change beliefs in competing hypotheses. Because of this, the empirical data that comes from modeling can be particularly informative about these aspects of acquisition, with respect to both what will (and what will not) work.

Acquiring a complex linguistic system is not an easy task because there is often a non-transparent relationship between the observable data and the underlying system (or grammar) that speakers used to generate those data. Children must effectively reverse engineer the data, inferring the pieces of the underlying system that combined to create them. As an example, consider the metrical phonology system, which determines the stress contour associated with a word. This includes which syllables are stressed, as well as how much stress syllables receive relative to each other. Here, we will consider only the basic division into stressed and unstressed. Suppose a child encounters the word *octopus*, which has the stress contour [stressed unstressed unstressed][1]. There are a variety of hypotheses a child might posit that are compatible with this stress contour, some of which are shown in (1). Perhaps all words ending with the sound *s* are stress-initial (1a), or all 3-syllable words have this stress contour (1b), or a word's stress contour results from dividing the word into units larger than syllables according to a number of options, and then stressing the leftmost syllable in each larger unit (1c).

* Many thanks to Amy Weinberg, Bill Idsardi, Jeffrey Lidz, Charles Yang, Roger Levy, and the audiences at GALANA 2008, the UCSD Linguistics Department, the UCI Artificial Intelligence and Machine Learning Group, the UCLA Linguistics Department, and the USC Linguistics Department for many wonderful comments.

[1] Stressed syllables will be indicated by underlining henceforth.

To appear in *Proceedings of GALANA 3*, Cascadilla Press.
Completed December 15, 2008

(1) Some hypotheses compatible with *octopus*

(a) Stress the first syllable only if the word ends with *s*

octopus ends with *s*: *octopus*

(b) Stress the first syllable only of 3-syllable words

octopus is a 3-syllable word: *octopus*

(c) Divide the word into units; stress the leftmost syllable in each unit

octopus: (*oc* *to*) *pus* = *octopus*

It is difficult to choose among these (and many other) hypotheses, given no other data or indications of which hypotheses are more likely. Here, we examine what children need in order to converge on the correct hypothesis for a realistic acquisition case study, English metrical phonology.

The remainder of the paper proceeds as follows. First, we briefly highlight general problems associated with acquiring complex linguistic systems, and discuss the utility of linguistic parameters and constraints. We then turn to the particular case study of the metrical phonology system, reviewing the specific instantiation considered here. Following this, we describe the specifics of the English system and the nature of the data available to English children. We then present several unbiased probabilistic learning models, and find that they all fail to converge on the correct grammar with any reliability. We subsequently identify the source of their failure and find that it is likely to cause *any* unbiased probabilistic learning model to fail. This suggests that children are not unbiased probabilistic learners. We then explore potentially helpful biases children might use during acquisition, and conclude with discussion of implications for acquisition.

2. Infinite Possibility and the Nature of the Problem

If we consider the metrical phonology system again, it is easy to see that acquisition of the correct system may be difficult from the available data, due to the number of variables that must be considered. For example, considerations may include if stress depends on the specific structure of the syllables involves, if units larger/smaller than syllables are formed, and how stress is assigned within these other units if such units are indeed formed. Given no limitations on what variables are relevant, children's hypothesis space of possible systems becomes infinite as it includes all possible variables in their linguistic (and non-linguistic) experience.

Restricting children's consideration to linguistic variables most likely to influence a stress contour lessens the problem, but does not solve it. For example, suppose children know that units larger than syllables are formed, which is fairly informative knowledge. Even then, they must consider variables such as (a) if all syllables in the word are included in the larger units, (b) which syllables within a larger unit are stressed, and (c) if rhyming matters in the formation of the larger units. Some of these variables will turn out to be relevant (e.g. (a) and (b)), but some will not (e.g. (c)). Each variable can then expand quite extensively if there are no restrictions. For example, if we focus on the variable regarding which syllables are included in larger units, options include (a) all syllables, (b) all but the leftmost syllable, (c) all but the rightmost syllable, (d) all but the leftmost and the rightmost syllable, (e) all but the leftmost syllable or all but the rightmost syllable, and so on. Without some constraints, the hypothesis space can again expand a great deal.

Part of what theoretical research offers is a way to constrain children's hypotheses to a substantially smaller part of the infinite hypothesis space. One idea for the appropriate hypothesis space comes from comparative linguistic research, where languages are observed to vary only in constrained ways from each other according to a number of variables. Children are then supposed to consider only hypotheses that are possible adult systems, as specified by these variables. The relevant variables are often a number of linguistic parameters (e.g. Chomsky, 1981; Halle & Vergnaud, 1987) or constraints (Tesar & Smolensky, 2000) – both serve the same purpose of circumscribing children's hypothesis space so it is no longer infinite.

A finite hypothesis space reduces the severity of the acquisition problem, but again does not solve it. For example, suppose there are n parameters with two options each. The child's hypothesis space then has 2^n grammars in it. Even if n is not very large, this can yield a very large hypothesis space. For example, if there are twenty parameters, there are 2^{20} = 1,048,576 grammars for children to choose

between. They can use the data to do this, but data are often ambiguous and so compatible with multiple hypotheses.

3. Case Study: Metrical Phonology

The metrical phonology system was chosen as a tractable case study since the hypothesis space can be explicitly defined by a reasonably small number of parameters. Note, however, that these parameters interact and make identifying the parameter value responsible for a given stress contour non-trivial (known as the Credit Problem (Dresher, 1999). We briefly review below the metrical phonology parameters considered here, adapted from Hayes (1995) and the parameterization in Dresher (1999) that draws from Halle & Vergnaud (1987). 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 work (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. It is unlikely children this age 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. Thus, this case study focuses on acquiring the first parts of the full metrical phonology system adult speakers use.

The first parameter, quantity sensitivity, concerns whether all syllables are identical, or differentiated by syllable rime weight. A language could be *quantity insensitive* (QI), so that syllables are undifferentiated. Multiple syllable types (short vowel with coda (VC), short vowel (V), and long vowel (VV)) are exemplified in *company* in (2), and all are represented by the undifferentiated syllable class 'S' in a QI analysis.

(2) QI analysis of *company*

syllable class	S	S	S
syllable rime	VC	V	VV
syllable structure	CVC	CV	CVV
syllables	com	pa	ny

A language could instead 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).

(3) QS-VC-L/H analysis of *company*

syllable class	H/L	L	H
syllable rime	VC	V	VV
syllable structure	CVC	CV	CVV
syllables	com	pa	ny

Stress assignment relies on both syllable weight and the formation of units larger than syllables called metrical feet.[2] For example, if a syllable is Heavy, it should be stressed. However, syllables not in metrical feet (*extrametrical* syllables) cannot be stressed; so, even if an extrametrical syllable is Heavy, it cannot receive stress. Languages with extrametricality (Em-Some) have either the leftmost syllable (Em-Left) or the rightmost syllable (Em-Right) not included in a metrical foot.[3] Example (4a) shows an Em-Left analysis of *giraffe* for a QS grammar; example (4b) shows an Em-Right analysis of

[2] Metrical feet will be indicated by parentheses (…) henceforth.

[3] Extrametrical syllables will be indicated by angle brackets <…> henceforth.

company for a QS grammar. Note in (4b) that the rightmost syllable, while Heavy, does not receive stress because of the parametric interaction with extrametricality.

(4a) Em-Some, Em-Left analysis of *giraffe*
syllable class <L> (H)
syllables gi raffe

(4b) Em-Some, Em-Right analysis of *company*
syllable class (H L) <H>
syllables com pa ny

In contrast, languages without extrametricality (Em-None) include all syllables in metrical feet. Example (5) below demonstrates an Em-None analysis for *afternoon*, which generates two metrical feet encompassing all the syllables in the word.

(5) Em-None analysis for *afternoon*
syllable class (L L) (H)
syllables af ter noon

Once the syllables to be included in metrical feet are known, metrical feet can be constructed. However, there is variation on which edge of the word metrical foot construction begins at. It can begin from either the left side (Ft-Dir-Left, (6a)) or the right side (Ft-Dir-Rt, (6b)).

(6a) Start metrical feet construction from the left (Ft-Dir-Left): (L L H
(6b) Start metrical feet construction from the right (Ft-Dir-Rt): L L H)

Then, the size of metrical feet must be determined. An *unbounded* (Unb) language has no arbitrary limit on foot size; a metrical foot is only closed upon encountering a Heavy syllable or the word's edge. Thus, there is a parametric interaction with the quantity sensitivity parameter value (if it is used in the language), which determines which syllables are Heavy. Also, there is a parametric interaction with feet directionality if a word contains Heavy syllables: starting metrical foot construction from the left (7a) can yield different metrical feet than starting from the right (7b). If there are no Heavy syllables or the syllables are undifferentiated, then the metrical foot encompasses all the non-extrametrical syllables in the word.

(7) QS Syllables, building metrical feet
(a) Ft-Dir-Left: (L L L) (H L)
(b) Ft-Dir-Rt: (L L L H) (L)

Another option is for metrical feet to be a specific, arbitrary size; these are *bounded* (B) languages. 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). Only if the word edge is reached can metrical feet deviate from this size. Examples (8-10) demonstrate different bounded analyses.

In (8), a B-2 analysis groups units by two, except for the final foot which is only one unit since the word edge is reached. If the counting units are syllables, it does not matter how the syllables are differentiated (9a-b), or even if they are differentiated (9c) – metrical feet will always contain the same number of syllables. However, if the counting units are moras (represented by μ in (10)), there is a parametric interaction with quantity sensitivity since Heavy syllables are two units while Light syllables are one. This can lead to different metrical feet than counting by syllables would – compare the B-Syl analysis in (9b) to the B-Mor analysis in (10).

(8) B-2 analysis over five units, Ft-Dir-Left
B-2, Ft Dir Left: (x x) (x x) (x)

(9) B-Syl analyses of a sequence of four syllables, B-2
(a) (L H) (L L)
(b) (H H) (L L)
(c) (S S) (S S)

(10) B-Mor analysis of a sequence of four syllables, B-2
mora analysis μ μ μ μ μ μ
syllable classification (H) (H) (L L)

Once the metrical feet are formed, one syllable per metrical foot is stressed. It can be either the leftmost syllable (Ft-Hd-Left, (11a)) or the rightmost syllable (Ft-Hd-Rt, (11b)).

(11) Feet Headedness analyses for the syllable sequence H L L divided into metrical feet
(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 the 156 grammars in the hypothesis space. As mentioned before, it may be quite difficult to determine if a particular parameter value is responsible for generating a particular stress contour because of parametric interaction. For example, consider two grammars that the word *cucumber* is compatible with (12). 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.

(12) 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

4. Case Study: 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 misleading data that favor the incorrect parameter values for English. For example, the English grammar is incompatible with approximately 27% of the available data (by tokens) – that is, for 27% of the data points, the child can only conclude that parameter values other than the English values are responsible for generating the data point. A reasonable question is if a parameterization that causes the English grammar to be incompatible with such a large portion of the English data is really a valid system to examine. While there obviously must be some way to deal with these 27% exceptional data, a grammar that can reliably cover 73% of the data is still a useful grammar for children to have. Moreover, many of these exceptions are due to interaction with the morphological system, and so 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. Third, previous computational modeling research (Pearl, 2008; Pearl, 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 acquisition success 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. 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, numerous English child-directed speech samples are available through CHILDES (MacWhinney, 2000), so realistic estimates of the data distributions children encounter can be made.

The input for the models 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 parameters of the metrical phonology system under consideration might be set. These corpora yielded 540505 words of orthographically transcribed child-directed speech. A child's syllabification of these words was estimated by referencing the MRC Psycholinguistic Database (Wilson, 1988), which gives adult syllabifications of English words. The stress contour a given word was pronounced with was estimated by both the MRC Psycholinguistic Database (Wilson, 1988) and the CALLHOME American English Lexicon (Canavan, Graff, & Zipperlen, 1997). A 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)).

5. Unbiased Models and Modeling Results

5.1. 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, submitted). The hypothesis space here is defined in terms of competing grammars, similar to other previous modeling work (Clark, 1992; Fodor & Sakas, 2004; Gibson & Wexler, 1994; Niyogi & Berwick, 1996; Pearl, 2008; Pearl & Weinberg, 2007; Sakas, 2003; Sakas & Fodor, 2001; Sakas & Nishimoto, 2002; Yang, 2002). 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, similar to several previous modeling studies (Fodor & Sakas, 2004; Gambell & Yang, 2006; Gibson & Wexler, 1994; Niyogi & Berwick, 1996; Pearl, 2008; Pearl & Lidz, submitted; Pearl & Weinberg, 2007; Sakas, 2003; Sakas & Fodor, 2001; Vallabha *et al.*, 2007; Yang, 2002). The motivation for using incremental/online models is that they are more likely to use algorithms that children use to acquire language, since children are not likely to have the memory capacity to store every utterance ever heard in all its detail for analysis later on. Instead, children are more likely to extract information from the data as the data are encountered.

5.2. Unbiased Models

The basic hypothesis space for each of the unbiased models considered is the set of 156 viable grammars discussed in section 3. For each parameter, there are two competing values (e.g. QS vs. QI for quantity sensitivity). The unbiased model initially associates a probability of 0.5 with each. This probability is then altered, based on the data encountered.

A given data point contains two types of information: the syllable structure (e.g. 'VV VC VC') and the stress contour (e.g. '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 (13a); if the generated stress contour does not match, all parameter values in that grammar are punished (13b). The model then moves on to the next data point.

(13) 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

no match: 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 assigning Em-Some 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 choose 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 (14). The update equation involves a parameter γ that determines how liberal the model is. The larger γ is, the more probability the model shifts for a single data point.

(14) Linear Reward-Penalty Scheme

p_i = previous probability of parameter value
p_k = previous probability of opposing parameter value

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

$$p_{i\text{-}new} = p_i + \gamma(1 - p_i)$$
$$p_{k\text{-}new} = 1 - p_{i\text{-}new}$$

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

$$p_{i\text{-}new} = (1 - \gamma)p_i$$
$$p_{k\text{-}new} = 1 - p_{i\text{-}new}$$

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 what 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 (15)). 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 value 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 are then normalized so they sum to 1.

(15) BayesLearner update equation

p_i = previous probability of parameter value

p_k = previous probability of opposing parameter value

$p_{i\text{-}new} = (\alpha + 1 + \text{successes})/(\alpha + \beta + 2 + \text{total data seen})$

$p_{i\text{-}new\text{-}normalized} = p_{i\text{-}new}/(p_{i\text{-}new} + p_k)$

$p_{k\text{-}new\text{-}normalized} = p_k/(p_{i\text{-}new} + p_k)$

A variation on these model types incorporates a method for smoothing the acquisition trajectory when the system to be acquired involves multiple parameters (Yang, 2002), such as the metrical phonology system here. This kind of model 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. This 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. We will refer to models of this kind as *count-learning* models.

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 $+/-c$. A count BayesLearner only updates if the counter reaches $+/-c$: 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.

5.3. Model Parameters and Simulation Results

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, and 0.025. The Count NPLearner and Count BayesLearner were run with count parameter $c = 2$, 5, 7, 10, 15, 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).

Table 1 shows the average percentage of the trials each model converged on the English grammar. The most striking aspect of these results is the extreme rarity with which these unbiased models 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. This is certainly not the robust convergence behavior we would expect from a model that accurately reflects the acquisition process children use. In short, the unbiased probabilistic models do not perform anywhere near as well as English children do.

Unbiased Model	Average % English Convergence
NPLearner γ = 0.001, 0.0025, 0.01, or 0.025	0.000
BayesLearner	0.000
Count NPLearner γ = 0.001, 0.0025, 0.01, or 0.025 *c* = 2, 5, 7, 10, 15, or 20	0.033
Count BayesLearner *c* = 2, 5, 7, 10, 15, or 20	0.000

Table 1. Unbiased Modeling Results

If we look closer at the modeling results 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 (16), we see several example grammars of this kind, with incorrect values in italics.

(16) 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*

6. Discussion

6.1. The Problem for Unbiased Models

We found exceedingly poor performance by unbiased probabilistic learning models, given realistic English data to learn from. What is the source of the problem? Consider the hypothesis space, which contains 156 grammars. In order for an unbiased model to converge on the English grammar, that grammar should be optimal (or near optimal) for the data the model is learning from. If we rank the competing grammars by their compatibility with the English data set, it turns out there are 51 other grammars 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: 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.

Since English children do presumably succeed given English input, one solution is to believe that children are not unbiased probabilistic learners. Instead, 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.

6.2. Bias on the Hypothesis Space

One bias English children may have relates to their hypothesis space. 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. *ba*ba) over weak-strong bisyllables (e.g. b*a*ba). 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 – only 60 grammars instead of 156 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. To simulate the learner's prior knowledge, the probabilities of QS and Ft-Hd-Left are initially 1.0 (rather than 0.5). Table 2 shows the average percentage of the trials each model converged on the English grammar. 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.

Model with Hypothesis Space Bias	Average % English Convergence
NPLearner γ = 0.001, 0.0025, 0.01, or 0.025	0.000
BayesLearner	0.100
Count NPLearner γ = 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 - 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.

6.3. 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 better than 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 set in particular orders. So, if the parameters are set in one of those orders, the English grammar is the optimal grammar for the unambiguous data. At that point, 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).

6.4. Implications for Acquisition

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 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 acquisition case studies (e.g. English anaphoric *one*: Pearl & Lidz (submitted), Regier & Gahl (2004); structure-dependence of syntactic rules: Perfors, Tenenbaum, & Regier (2006)). The present study reinforces this for acquiring parametric systems such as metrical phonology. The exact nature of the necessary bias can be investigated through computational modeling studies, such as Regier & Gahl (2004) which uses an implicit data intake bias and a subset bias on the hypothesis space, Perfors, Tenenbaum, & Regier (2006) which uses a simplicity bias on the hypothesis space, Pearl (2008) and Pearl (submitted) which use a data intake bias, and Pearl & Lidz (submitted) which also uses a data intake bias. Of particular interest is whether the necessary bias is likely to be domain-specific (Pearl, 2008; Pearl, submitted; Pearl & Lidz, submitted; Regier & Gahl, 2004) or domain-general (Pearl & Lidz, submitted; Perfors, Tenenbaum, & Regier, 2006; Regier & Gahl, 2004).

In general, complex linguistic systems – such as metrical phonology – will likely require more than probabilistic learning in order to acquire the knowledge children acquire, 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.

References

Akhtar, Nameera, Callanan, Maureen, Pullum, Geoffrey, & Scholz, Barbara. (2004). Learning antecedents for anaphoric *one*. *Cognition 93,* 141-145.

Bernstein Ratner, Nan. (1984). Patterns of vowel Modification in motherese. *Journal of Child Language 11*, 557-578.

Brent, Michael & Siskind, Jeffrey. (2001). The Role of Exposure to Isolated Words in Early Vocabulary Development. *Cognition 81/82*, 33–44.

Bush, R. & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review 58*, 313-323.

Canavan, Alexandra, Graff, David, and Zipperlen, George. (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.

Dresher, Elan. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry 30*, 27-67.

Fodor, Janet & Sakas, William. (2004). Evaluating Models of Parameter Setting. *Proceedings of the 28th Annual Boston University Conference on Language Development,* 1-27.

Gambell, Timothy & Yang, Charles. (2006). Word Segmentation: Quick but not dirty. Manuscript: Yale University.

Gerken, LouAnn & Aslin, Richard. (2005). Thirty years of research on infant speech perception: The legacy of Peter W. Jusczyk. *Language Learning and Development 1*, 5-21.

Gibson, Edward & Wexler, Kenneth. (1994). Triggers. *Linguistic Inquiry 25*, 407-454.

Gómez, Rebecca & Lakusta, Laura. (2004). A first step in form-based category abstraction by 12-month-old infants. *Development Science 7(5)*, 567-580.

Halle, Morris & Vergnaud, Jean-Roger. (1987). *An Essay on Stress*. Cambridge, MA: MIT Press.

Hart, Betty & Risley, Todd. (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.

Hudson Kam, C., & Newport, E. (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, Cutler, Anne, & Redanz, Nancy. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development 64*, 675-687.

MacWhinney, Brian. (2000). The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates.

Niyogi, Partha, & Berwick, Robert. (1996). A language learning model for finite parameter spaces. *Cognition 61*, 161-193.

Pearl, Lisa. (2008). Putting the Emphasis on Unambiguous: The Feasibility of Data Filtering for Learning English Metrical Phonology. *BUCLD 32: Proceedings of the 32nd Annual Boston Conference on Child Language Development,* 390-401.

Pearl, Lisa. (submitted). Acquiring Complex Linguistic Systems From Natural Language Data: What Selective Learning Biases Can Do. Manuscript: University of California, Irvine.

Pearl, Lisa & Lidz, Jeffrey. (submitted). When domain general learning fails and when it succeeds: Identifying the contribution of domain specificity. Manuscript: University of California, Irvine & University of Maryland, College Park.

Pearl, Lisa & Weinberg, Amy. (2007). Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling. *Language Learning and Development 3(1)*, 43-72.

Perfors, Amy, Tenenbaum, Joshua, & Regier, Terry. (2006). Poverty of the Stimulus? A rational approach. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Regier, Terry & Gahl, Susanne. (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 Fodor, Janet. (2001). The Structural Triggers Learner. In S. Bertolo (ed.), Language Acquisition and Learnability. Cambridge: Cambridge University Press.

Sakas, William & Nishimoto, Eiji. (2002). Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Manuscript: City University of New York.

Tesar, Bruce & Smolensky, Paul. (2000). Learnability in Optimality Theory. Cambridge, MA: The MIT Press.

Turk, Alice, Jusczyk, Peter, & Gerken, LouAnn. (1995). Do English-learning Infants Use Syllable Weight to Determine Stress? *Language and Speech 38(2)*, 143-158.

Vallabha, Gautam, McClelland, James, Pons, Ferran, Werker, Janet, & Amano, Shigeaki. (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.