Running Head:

Parametric system acquisition with selective learning biases

Article Type:

Full Article

Title:

Acquiring parametric linguistic systems from natural language data: What selective learning biases can do

Author, Affiliation, & Contact Information:

Lisa Pearl

University of California, Irvine

Department of Cognitive Sciences

3151 Social Science Plaza

University of California

Irvine, CA 92697-5100

Phone: 949-824-0156

Fax: 949-824-2307

Email: lpearl@uci.edu

Acquiring parametric linguistic systems from natural language data:

What selective learning biases can do

Abstract

Parametric systems have been proposed as models of children's knowledge representations about language, often as a way for them to acquire the specific linguistic knowledge they eventually attain. A reasonable test, then, is to see if children could in fact acquire the proposed parametric systems from the data available to them. One might think it necessary to use all available information since natural language data are often ambiguous and noisy. The case study here suggests that having a selective learning bias to learn only from unambiguous data leads to acquisition success for an instantiation of the English metrical phonology system involving nine parameters. Special attention is given to the model's input and the psychological plausibility of the knowledge states, learning biases, and algorithms used in the model, in order to consider the learning problem from the perspective of children acquiring the linguistic systems of their native language. The results support both the unambiguous data bias as a viable strategy for acquisition, and also the parametric instantiation of the metrical phonology system since the correct system can be learned from the natural language data children encounter.

key words: acquirability, acquisition, English, metrical phonology, natural language data, parametric systems, selective learning bias, unambiguous data

1. Introduction

1.1. Language learnability and language acquisition

Knowledge of language consists of many kinds of systematic knowledge, such as phonology, morphology, and syntax. The task of language learners is to uncover this systematic knowledge for their native language. If the learner is a child, this process is often termed "acquisition", and there are several constraints on the learnability scenario, including the type of input the learner receives, how long the learner has to learn certain knowledge, and how the learner is able to process the input. For this reason, as Johnson (2004) notes, "acquirability" is somewhat different from the learnability traditionally considered in computational learning literature (e.g. Gold (1967)), "even when children are identified with learning functions and natural languages with sets of sentences", as might be the case for syntactic acquisition. Often, the conflation of learnability with acquirability has led linguists and psychologists to misinterpret the implications of computational learnability results for child language acquisition (e.g. see Johnson (2004) for a review of the numerous misinterpretations of Gold's (1967) result). Here, we consider the acquirability of a parametric linguistic system from natural language data, keeping the restrictions that come with acquisition in mind. In this way, the learning results from this study can be more directly transferred to acquisition research.

1.2. The tricky business of acquisition for complex 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 in the linguistics literature. 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 acquiring a specific language converge quickly on the complex 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: 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.

It therefore seems reasonable to ask if a given proposal for prior knowledge makes the complex system it is designed to help acquire actually acquirable. One proposal is a parametric system, and one domain for which it has been proposed is metrical phonology (Halle & Vergnaud (1987), Hayes (1995)). This study examines the acquirability of a parametric system of metrical phonology for English from the input available to English children. If this system is not acquirable from realistic data, then we have mark against that proposal. If instead this system is acquirable, it is viable as a proposal of the knowledge representation in children's minds, and we can then explore the conditions under which it is acquirable.

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 parameters 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 of 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 (sometimes referred to as 'grammars'). In addition, knowing what 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).

In addition, the known 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)	(b)	(c)
(S S) (S)	(L L) (H)	(L) (L H)
af ter noon	af ter noon	af ter noon

Metrical phonology system parameters include which syllables are included 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 in fact solved. The acquirability of the correct grammar from the available data is still an open question.

1.3. A framework for acquisition and learning biases

If we consider the language acquisition mechanism, at least three separate pieces can be identified: the hypothesis space, the data intake, and the update procedure (Pearl 2007). The hypothesis space consists of all the hypotheses currently under consideration. For instance, this could be the set of potential grammars available from the combination of the different parameter values available. The data intake refers to the set of data children learn from (Fodor 1998b). This may be the entire input set, or some subset of it. The update procedure specifies how children change belief in the various competing hypotheses, based on their data intake. This procedure is often instantiated in acquisition models as a domain-general learning algorithm that shifts probability among the hypotheses (e.g. Bayesian learning: Tenenbaum & Griffiths (2001), Perfors *et al.* (2006), Foraker *et al.* (2007); Linear reward-penalty: Yang (2002)).

With respect to acquisition, there are potentially advantageous constraints that can be placed on different pieces of this mechanism, which might be termed "learning biases". A bias on the hypothesis space is knowing the relevant parameters and their respective

potential values, thereby restricting the hypothesis space to a subset of what it would otherwise be. As discussed above, this bias serves to make the correct grammar more likely to be acquired.

Another learning bias children might use relates to the data intake. Specifically, children might selectively learn only from data they perceive as maximally informative: *unambiguous* data (Fodor 1998a, Dresher 1999, Lightfoot 1999, Pearl & Weinberg 2007). This unambiguous data bias would effectively be a data intake filter implemented by the child's acquisition mechanism. The filter then causes the child to ignore information in ambiguous data, and focus instead on information available in the data perceived as unambiguous in order to identify the correct grammar.

While learning only from maximally informative data has intuitive appeal, it is not without its difficulties. As mentioned above, data are often ambiguous, especially in systems involving multiple interacting parameters, such as metrical phonology. So, unambiguous data would comprise only a small subset of the available input, if such data exist at all (Clark 1994). A reasonable concern is the viability of this kind of selective learning bias, given a realistic parametric system to learn and realistic data to learn from. In short, though unambiguous data are highly informative, do they exist in the natural language data children encounter? If they do exist, do they exist in sufficient quantities to lead children to the correct grammar?

For any given acquisition scenario, an unambiguous data bias may prove detrimental if the answer to either of these questions is no. The existence of unambiguous data is an empirical question that must be examined for a particular acquisition problem, as is the

existence of sufficient quantities for learning the correct grammar. The identification of unambiguous data is a question that has been considered in various domains (e.g. syntax: Fodor 1998a, Lightfoot 1999, Pearl & Weinberg 2007; metrical phonology: Dresher 1999), and promising proposals can be tested against a particular acquisition problem to assess their viability.

1.4. The present study: Realistic acquisition scenarios

Here we examine the viability of an unambiguous data bias for the acquisition of a parametric system of English metrical phonology. The system we consider includes nine parameters (adapted from Dresher (1999) and Hayes (1995)). The data available to children (estimated from the CHILDES database (MacWhinney 2000)) are quite ambiguous and contain many exceptions to the English grammar. Both the complexity of the system and the noisiness of the data make converging on the correct grammar for English a non-trivial acquisition problem. The existence of unambiguous data for the parameters is not assured; the existence of unambiguous data in sufficient quantities to lead the child to English is definitely not assured.

Previous computational work on parametric metrical phonology systems, while exploring systems of similar complexity, has not used child-directed speech as input (Dresher & Kaye 1990, Dresher 1999) when testing the system's acquirability. To address this, the model here uses a data set as input that contains both the forms children are likely to encounter and the frequencies at which they will encounter these forms. Note that these data differ from adult-directed speech in several respects, so the use of

child-directed speech is important for testing acquirability. See the appendix for a detailed comparison of the child-directed speech used here and adult-directed speech.

Because we intend to measure the performance of an unambiguous data bias on the specific case study of English metrical phonology, we will briefly review the particular parametric system under consideration, the values for the target language English, and the distributions from English child-directed speech that are used as input. We will then describe two classes of proposals (*cues*: Dresher 1999, Lightfoot 1999; *parsing*: Fodor 1998a, Sakas & Fodor 2001) for how children can identify unambiguous data in their input. These proposals will then each be used to implement an unambiguous data filter for the scenario where the child attempts to acquire English metrical phonology.

We will see that a probabilistic learning model using an unambiguous data filter implemented with either of the two identification methods can in fact converge on English under certain conditions. This is true despite highly ambiguous and exception-filled data. However, each identification method requires some additional knowledge in order to converge on English. The knowledge necessary for acquisition success will be discussed in each case, as well as the predictions that are generated from this required knowledge. We conclude with some general remarks on theoretical claims about how knowledge is represented in children's minds.

2. The parametric system of metrical phonology

The instantiation of the metrical phonology system considered here has nine parameters (five main parameters and four sub-parameters), adapted from the systems

described in Dresher (1999) and Hayes (1995). This system concerns only whether syllables are stressed or unstressed, and not how much stress syllables receive compared to other syllables. Moreover, this system does not include interactions with the English morphology system, though such interactions are thought to be fairly pervasive in English (see Chomsky & Halle (1968), Kiparsky (1979), and Hayes (1982) for several examples). This is due to considerations of the child's likely initial knowledge state when acquiring the metrical phonology system. Experimental work (Jusczyk *et al.* 1993, Turk *et al.* 1995) has suggested that children under a year old may already be acquiring some aspects of the English metrical phonology system. Kehoe (1998) suggests that children already know several parameter values of the English system by 22 months. It is unlikely that children of this age have extensive knowledge of English's morphology system, and so they may not hypothesize the interaction between the morphology system and the metrical phonology system in English.

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 the full system that will account for all of English, is still significantly more complex than parametric systems explored in some prior computational modeling work which involved

at most three interacting parameters (Gibson & Wexler 1994, Niyogi & Berwick 1996, Pearl & Weinberg 2007). Previous work that has examined parametric systems of equal or greater complexity has often not been empirically grounded with realistic input distributions (Dresher 1999, Sakas & Nishimoto 2002, Sakas 2003, Fodor & Sakas 2004, among others). In addition, a system very similar to the one here has been used to study the acquisition of stress in English as a second language (Archibald 1992).

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 included 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>
el	e	phant

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, refers to whether all syllables are identical in the system, or differentiated by syllable rime weight (Hayes 1980, Halle & Idsardi 1995, Dresher 1999, among many others). 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 *company*

QS analysis	L/H	L	H	QI analysis	S	S	S
syllable rime	VC	V	VV				
syllable structure	CVC	CV	CCVV				
syllables	com	pa	ny		com	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 included 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	gi raffe		com pa	ny

(4b) An Em-None analysis

syllable class	(L L)	(H)	
syllable rime	VC	VC	VV
syllables	af	ter	noon

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

(S̲ S) (S̲) <S> → S̲ S S̲ S

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

<S> (S̲ S) (S̲) → S S̲ S S̲

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.

3. English

Previous computational models that explored the acquirability of parametric metrical phonology systems (Dresher & Kaye 1990, Dresher 1999) have not used realistic estimates of the data that children are likely to encounter. So, while these systems may be learnable given certain data, it is unclear if they are acquirable given the data that children have access to. Perhaps these methods only work when the data set contains very long words or words in certain frequencies. The model presented here attempts to address this by using child-directed speech to estimate the model's input.

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, and with approximately 37% by types – 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. Hochberg (1988) finds that children's "propensity to hypothesize linguistic rules is so strong as to tolerate a high degree of exceptionality", suggesting that children may still search for an underlying system even though there are numerous exceptions. Also, this situation is not too unusual for metrical acquisition data; for example, Daelemans *et al.* (1994) note that 20% of the Dutch data they consider are irregular according to a generally accepted metrical analysis and so must be dealt with in terms of idiosyncratic marking. Another way of framing this situation is that the regular data compatible with the English grammar are the core data,

and the exceptional/irregular data are the periphery data that must be accounted for by some other means (e.g. Daelemans *et al*. (1994) suggest exception features, or associating the irregular pattern with the specific lexical item). 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 the English grammar in the data.

Another reason for choosing English is that previous computational modeling research (Pearl, to appear) has found that unbiased probabilistic models are unable to acquire the English grammar from child-directed English speech, and concluded that some kind of bias is required if children are to accomplish this. This paper provides an exploration of one plausible learning bias, which is to learn only from data perceived as unambiguous. The final reason for choosing English is that numerous English child-directed speech samples are available through CHILDES (MacWhinney 2000), so realistic estimates of the data distributions children encounter can be obtained.

The Bernstein-Ratner corpus (Bernstein 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, given that several parameters of this system seem to be known by 22 months (Kehoe 1998). In total, this yielded 540505 words of orthographically transcribed child-directed speech, consisting of 8093 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.

4. Unambiguous data

The acquisition problem for this case study is fairly difficult since the child must successfully navigate the vagaries of the data in order to converge on the correct parameter values for English. It is possible that a useful bias for a child to have is to learn only from data perceived as unambiguous. But, as noted earlier, many data are ambiguous. The data perceived as unambiguous therefore comprise only a small subset of the available input, if they exist at all. This makes learning only from unambiguous data a potentially dangerous strategy. Moreover, there is no guarantee that unambiguous data will appear in the correct relative quantities to converge on the English values. For instance, even if unambiguous data exist for having some extrametricality (Em-Some), *more* unambiguous data may exist for having no extrametricality (Em-None). As Em-Some is the correct value for English, this acquisition scenario is problematic for an English child even in the case that unambiguous data do exist.

In addition, the proposals that describe how a child could identify unambiguous data (*cues* (Dresher 1999) and *parsing* (Fodor 1998a, Sakas & Fodor 2001)) add in extra variation. In each proposal, the way a child identifies any given data point as unambiguous for some parameter value can depend on the child's current knowledge about the system as a whole (a property sometimes called "progressive disambiguation of the input" (Sakas 2000) or "dynamic disambiguation" (Sakas & Fodor 2001)). So, what is perceived as unambiguous can change as the child acquires more knowledge about the system. Data that are ambiguous early on in the acquisition process may be viewed as unambiguous later on once the child knows more of the target language's parameter values; data unambiguous initially may later be viewed as exceptional if they don't accord with the parameter values then known. A data point's status as unambiguous will be gauged subjectively by the child, and so will change over time. This represents the idea that the information an unambiguous learner garners from the data depends on what the learner already knows. This seems a desirable property from an information-theoretic standpoint (Shannon 1948), but it does make "unambiguous data" a moving target since a data point's status can change over time.

In the next section, we review the two proposals for identifying unambiguous data, and define how they work for acquiring the parametric metrical phonology system.

4.1 Identification via cues

A cue is a "specific configuration in the input" associated with a particular parameter value (Dresher 1999). The cues presented here match the observable form of a data point

– in this case, the combination of syllable structure and stress. The presence of a cue signals to the child that one parameter value is preferred over another for a given parameter. Cues for each value of the metrical phonology system are given in Table 1, with an example of each cue in parentheses after the description of the cue. Note that most cues depend on the current state of the child's knowledge (e.g. see the cues for QS, Ft-Dir-Left, B-Syl, and Ft-Hd-Left).

It should also be noted that the cues advocated here are not the cues proposed in Dresher (1999), but are designed in the same spirit – to identify highly informative data. Unlike some of the cues in Dresher's proposal, all the cues here can be identified within a single data point. This is in contrast to cues that operate over multiple data points. Not needing to compare multiple data points may be desirable if the child is simply extracting information from the current data point and integrating that information into her knowledge of the parametric system, rather than explicitly comparing the current data point to items already in the lexicon. In addition, cues are proposed not just for those parameter values that could be viewed as marked, but also for parameter values that could be viewed as the default option.

[Put table 1 approximately here: *Cues for metrical phonology parameter values.*]

4.2 Identification via parsing

The parsing method involves the child using the structure-assigning ability of parsing that is presumably used already during language comprehension (Fodor 1998a, b, Sakas

& Fodor 2001). The parsing instantiation we examine here tries to analyze a data point with "all possible parameter value combinations", conducting an exhaustive search of "all parametric possibilities" (Fodor 1998a). We will call this the *find-all-parses* approach, though it has also been called the Strong Structural Triggers Learner approach (Sakas & Fodor 2001, Fodor & Sakas 2004). Note that there are numerous implementations of parsing that could be tried (see Fodor & Sakas (2004) for a review), often with different strengths and weaknesses than the implementation examined here. The find-all-parses variant was chosen for this study as it contrasts most strongly with the cues method, as will be discussed below.

For *find-all-parses* parsing, a successful parameter value combination will generate a stress contour that matches the observed stress contour of the data point - this is then a successful parse of the data point. For instance, the combination (QI, Em-None, Ft-Dir-Left, B, B-2, B-Syl, Ft Hd Left) is able to generate the stress contour [stressed unstressed stressed] for the word *afternoon*. Since the stress contour the child would encounter for *afternoon* matches this stress contour (*afternoon*), this combination can successfully parse this data point.

If all successful parses use only one of the available parameter values for a given parameter (e.g. Em-None of the extrametricality values), that data point is viewed as unambiguous for that parameter value. Data points that can be parsed with multiple parameter values of the same parameter (e.g. Ft-Hd-Left and Ft-Hd-Rt for the feet headedness parameter) are considered ambiguous. These ambiguous data points are

filtered out of the child's intake for that parameter value (e.g. feet headedness) by the child's unambiguous data learning bias.

As an example of this parsing method in action, suppose the child encounters *afternoon*, and successfully recognizes two pieces of information: (1) the syllables are *af* (VC), *ter* (VC), and *noon* (VV), and (2) the associated stress contour is VC VC VV. A find-all-parses child would try to generate the observed stress contour with all available parameter value combinations and come up with five that are successful (10). Note that a parameter value ceases to be available when the child has converged on the opposing parameter value for the language. For example, if the child has decided the language's metrical feet are QS, then the QI value will no longer be available. So, at that point, the child will only try parameter value combinations using the QS value.

All the successful parses in (10) share Em-None, meaning that Em-None was required for a successful parse. The child then perceives this data point as unambiguous for Em-None.

(10) Successful parameter value combinations for *afternoon*: Em-None required

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

(b) (QI, **Em-None**, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Rt)

(c) (QS, QS-VC-L, **Em-None**, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Left)

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

(e) (QS, QS-VC-L, **Em-None**, Ft-Dir-Left, Unb, Ft-Hd-Left)

Recall that the informativity of a data point changes over time; more specifically, what a data point may be unambiguous for changes depending on the child's current knowledge of the adult language. We observed this in the cues learner as a change in what the cues for parameter values look like in the observable data (see table 1 where, for example, the cue for QS shifts based on the child's knowledge of extrametricality). This same malleability occurs for a parsing learner. Taking the data point from (10), we saw that five successful parses exist if all parameter values are available. However, suppose the child has some knowledge of the target language, specifically that English is bounded (B). Then, the previously successful parse using the unbounded (Unb) parameter value (10e) will no longer be tried, since it uses the incorrect parameter value. The remaining successful parses have more in common than Em-None: they share all of the Bounded values as well: B, B-2, and B-Syl (11). So, this very same data point will now be viewed by the child as unambiguous for Em-None, B, B-2, and B-Syl.

(11) Successful parameter value combinations for *afternoon*: B known

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

(b) (QI, **Em-None**, Ft Dir Rt, **B, B-2, B-Syl**, Ft-Hd-Rt)

(c) (QS, QS-VC-L, **Em-None**, Ft-Dir-Left, **B, B-2, B-Syl**, Ft-Hd-Left)

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

4.3 Cues and parsing: A quick comparison

To identify unambiguous data with cues, a child needs only to match the cue to the observable data – this makes identification simple. In addition, a cue can match a subpart

of the data point and so, for example, can match a portion of a word. This means the child can glean information without understanding the structure for the entire data point. This is also advantageous if the remaining portion of the data point is exceptional in some fashion (for instance, an unusual stress contour resulting from emotional speech). Because cue learners can extract information from partial comprehension, cues offer a way to get off the ground when children do not know much about the system.

Another advantage of learning with cues is the ability to easily incorporate default parameter values. For some parameters, a default assumption about the parameter value may be quite natural – e.g. since the child is initially dividing the word into syllables, using syllables as the counting unit for metrical feet (B-Syl) might be a more natural initial hypothesis than counting by moras (B-Mor). A child using default values for parameters would assume the default value is true unless there is evidence from the input to the contrary. The cue child can collect evidence for the non-default value by using that value's cue, which is quite important if the adult system actually uses the non-default value. However, if the language uses the default value, the child's work is partially done already – the child does not need to explicitly learn this value from the input.

Still, the main pitfall of cues is that the child must already have knowledge of what the cues are in order to learn this way. In addition, cues are heuristic, and may lead to false positives or false negatives that could have a detrimental effect on acquisition. In (12), we see an example of both error types. In the false positive example (12a), a cue for metrical feet headed on the left (Ft-Hd-Left) can match a data point that was generated using the feet headed right (Ft-Hd-Rt) parameter value. In the false negative example

(12b), the cue for quantity sensitivity (QS) does not match a data point that was generated using the QS parameter value.

(12) The heuristic nature of cues

(a) False positive: Ft-Hd-Left cue matches a data point generated with Ft-Hd-Rt

Ft-Hd-Left cue: Leftmost syllable is stressed

Ft-Hd-Rt structure: (L) (L H)

af ter noon

(b) False negative: QS cue misses a data point generated with QS

QS cue (Em-None/unknown): 2 syllable word with 2 stresses

QS structure: (L) (L H)

af ter noon

Turning to a find-all-parses learner, we find a rather complementary array of strengths and weaknesses.[1] We begin with the weaknesses. First, identification of unambiguous data is a non-trivial process, requiring the child to find all parameter value combinations that can parse the given data point. Second, the entire data point must be parsed in order for any information to be extracted. If exceptions exist in one portion of the data point, no information from the data point can be used since it cannot be parsed. This can make the initial stages of acquisition quite difficult, when the learner may not know enough to successfully analyze the entire data point.[2]

Third, no matter what the instantiation of parsing, a parsing learner cannot have a default value for a parameter and still collect data for the opposing parameter value. Specifically in that scenario, the only values available to the parser would be the default values. The parsing method cannot comprehend data that are unambiguous for the non-default values since it cannot parse such data with the default values. This causes the child to be unable to recognize unambiguous data for the non-default values, a problem noted by Valian (1990). So, the learning benefit gained from default values is unavailable to a parsing learner.[3]

Still, the main strength of parsing is that it does not require any additional knowledge beyond the ability to parse. Unlike the cues method, the child does not need any knowledge beyond the parameters themselves. In addition, find-all-parses parsing is not heuristic, and will only identify data that are truly unambiguous. So, a find-all-parses child will not be led astray by false positives or negatives.

To sum up, cues and find-all-parses parsing are two methods a child could use to identify unambiguous data in the input. Both methods have strengths and weaknesses, many in fact complementary. However, despite their weaknesses, cues and parsing share a major strength: both are compatible with incremental learning models in the sense that they extract information from a data point as it comes in.

The compatibility of these learning methods with incremental learning is an important point, as acquisition models should consider the psychological plausibility of the methods they employ (noted by Vallabha *et al.* (2007)). In the case of incremental learning, young

children have limited memories, so it is important to consider that they probably cannot hold large quantities of data in mind for analysis. This is not to say that they must instantly forget data that they hear; rather, it is that they cannot perfectly recall every detail of every data point they hear. So, they are likely to extract the relevant information from the current data point and incorporate it into their current knowledge state – in the case of metrical phonology acquisition, knowledge about the word type and about the metrical phonology system. This contrasts with batch-learning, where the model may operate over the entire corpus's worth of data (e.g. Perfors *et al*. 2006, Foraker *et al*. 2007, Goldwater *et al*. 2007, Hayes & Wilson, 2008). It seems reasonable for a model of acquirability to process data incrementally, though previously integrated information from prior data could be available (e.g. in the probabilities assigned to competing parameter values (Yang 2002)).

5. Selective learning

We turn now to the implementation of the probabilistic learning procedure a child would follow to acquire English metrical phonology. The main idea is fairly straightforward: the child encounters a data point; if it is perceived as unambiguous for any parameter values, the probability of those parameter values is updated (either increased or decreased as appropriate). The key intuition we will use is the following: if a child is trying to set a given parameter (P), the parameter value (P1 or P2) that has more unambiguous data in the input will eventually win the probabilistic learning race. Note that this will not apply for all learning algorithms. More specifically, while batch-learning

algorithms can have this property, not all algorithms with this property are batch-learning (e.g. see the Naïve Parameter Learner of Yang (2002) which relies on the frequency of unambiguous data, and Yang (2004) for empirical support of the relation of this frequency to children's trajectory of acquisition). For algorithms that have this property we can explore their behavior without explicitly implementing every learning algorithm of this kind. We note also that models could be setting parameters simultaneously, but it is the frequency of the unambiguous data that determines which parameters are set when (Yang 2002, Yang 2004). Crucially, the updating algorithm is only deployed for unambiguous data, rather than for ambiguous data. So, it does not matter how much ambiguous data the child encounters in the input. All that matters is the relative frequency of the unambiguous data for a given parameter's values. In addition, more than one unambiguous data point appears to be required to set a parameter, perhaps because children are aware that the world is a noisy place and they should be careful about making too big a change to their hypothesis without consistent support for that change. Instead, gradual change and acquisition based on the frequency of unambiguous data appears to be what happens in several cases (Yang 2004).

As long as the probabilistic learning procedure children use has this property of choosing the most probable value based on the available unambiguous data, we can predict what parameter values children will converge on by examining the input data. This allows us to be fairly agnostic about the particular details of the learning algorithm – for instance, it could easily be some instantiation of a linear reward-penalty scheme (Yang 2002), Bayesian learning (Tenenbaum & Griffiths 2001), online Expectation-

Maximization (Vallabha *et al.* 2007), or some other algorithm. The key property is that the incremental algorithm chooses the more probable value, based on the unambiguous data it observes over time. So, we can predict the parameter value children will converge on in the following way: the parameter value whose unambiguous data have a higher probability in the intake set will be the value the child converges on over time.

5.1. A learning procedure example

As a concrete example based on the Naïve Parameter Learner algorithm of Yang (2002), suppose the child is trying to determine whether the language has extrametricality (Em-Some vs. Em-None). The child will encounter a series of data points from the input, one at a time. Many of these data points will be ambiguous, but some will be unambiguous for Em-Some (say, 2.4% of them) and some for Em-None (say, 4.8% of them). Every time an Em-Some data point is encountered, the probability of Em-Some is increased some amount while Em-None is decreased (say, .01); every time an Em-None data point is encountered, the reverse is true. Suppose the child initially assigns equal probability to Em-None and Em-Some (.50 each). After an Em-Some data point, probability is shifted, and p(Em-None) = 0.49 while p(Em-Some) = 0.51. After an Em-None data point, p(Em-None) = 0.50 while p(Em-Some) = 0.50. After another Em-None data point, p(Em-None) = 0.51 while p(Em-Some) = 0.49. After 48 Em-None data points and 24 Em-Some data points (the number we would expect to find in 1000 input data points), p(Em-None) = 0.75 and p(Em-Some) = 0.25. So, the probability of Em-None steadily increases, due to the higher probability of its unambiguous data points. Given

enough input, the child will eventually converge on Em-None (usually after crossing some threshold deemed close enough to 1.0). Thus, in this scenario, the difference in the probability of encountering unambiguous data for the two parameter values is what determines the winning value for this incremental algorithm.

5.2. An example of the "moving target" status of unambiguous data

We have seen that a data point's status as unambiguous depends on what the child already knows about the metrical phonology system. So, the parameters that are set influence the data the child perceives as unambiguous for the unset parameters. This means that the probabilities of the unambiguous data that the child perceives can change as the child learns more of the English parameter values. The initial probabilities are not necessarily the same as the probabilities after the first parameter value is set; those are not necessarily the same as the probabilities after the second parameter value is set, and so on. The way the probabilities will change depends on which parameter value is set. The order in which parameters are set will thus determine the unambiguous data probabilities at any given point in the acquisition of the metrical phonology system. These probabilities may either favor or disfavor the correct parameter value for a given parameter, depending on what parameter values are set previously. So, the order in which parameters are set may determine if they are in fact set correctly, an idea noted in Dresher (1999).

As an example, consider Tables 2 and 3, which show the probability of encountering unambiguous data for each available parameter value at different points during

acquisition for a find-all-parses child. Each probability represents the likelihood that a given data point the child encounters will be perceived as unambiguous for a given parameter value. These are estimated by calculating the quantity of unambiguous data points in the available corpus given the child's current knowledge state (e.g. 2151 tokens) and dividing by the total number of data points in the available corpus (540505 tokens).

Table 2 shows the probabilities before any parameters are set, while table 3 shows the probabilities after QS is set. In both tables, the probabilities are quite small, since much of the input is ambiguous when the child is in either of these knowledge states (i.e. initially having no parameters set, then knowing only that the language is quantity sensitive). Still, for most parameter values, some unambiguous data does exist initially. This answers the first question of unambiguous data existence for this particular acquisition sc scenario – unambiguous data do indeed exist (or at least, data perceived as unambiguous exist). The second question of existence is whether unambiguous data exist in the correct relative quantities – that is, do the unambiguous data probabilities favor the correct parameter value for English? If we look at the probabilities in table 2, we see that initially the unambiguous data probabilities do favor the correct value for some parameters: QS, Ft-Dir-Rt, B, and Ft-Hd-Left. However, they do not favor the correct value for extrametricality, Em-Some. After QS is set however, the child will perceive different data as unambiguous (Table 3). Happily for the English child, the Em-Some and Em-None probabilities have changed so that Em-Some is now favored. So, the correct extrametricality value can indeed be learned from the data, but only if a certain parameter-setting order is obeyed. In this case, QS must be set before Em-Some.

Note that setting a parameter can have two effects: (1) altering the probabilities of encountering unambiguous data for other parameter values, and (2) opening up sub-parameters in the system. Before QS is set in Table 2, p(Em-None) is much higher than p(Em-Some) (.0284 vs. 0.0000259). After QS is set in Table 3, the reverse is true: p(Em-Some) is more than twice p(Em-None) (.0485 vs. .0240). In addition, a sub-parameter under quantity sensitivity is now available for how the child should treat VC syllables (QS-VC-L vs. QS-VC-H). Before the child has set the QS value, the sub-parameter is not relevant; after the QS value is set, the sub-parameter is relevant and so the child will start learning from the unambiguous data for those sub-parameter values.

[Put Table 2 approximately here: *Initial probabilities of unambiguous data.*]
[Put Table 3 approximately here: *Probabilities of unambiguous data after QS is set.*]

We can thus see explicitly how the order of parameter-setting can matter – setting one parameter value can easily influence the setting of subsequent parameters. From the example in tables 2 and 3, we see that an English child who sets quantity sensitivity to QS will subsequently set extrametricality to Em-Some. Conversely, an English child who sets extrametricality first will choose Em-None, which is incorrect for English.

5.3. Parameter-setting orders: potential solutions

Given the difficulty of this acquisition problem, success is by no means guaranteed. The worst case is that learning from unambiguous data is not viable: no parameter-setting

order will allow the child to converge on the English grammar. An unambiguous data learning bias is no help as unambiguous data for each parameter do not exist in the correct relative quantities at any point during acquisition. In a better case, learning from unambiguous data is viable, as there are parameter-setting orders available that will lead to English. As long as a probabilistic learning child sets the parameters in a viable parameter-setting order, that child will converge on the English parameter values.

To determine which (if any) parameter-setting orders lead to the English grammar, an exhaustive search was conducted of all 24,943,680 possible parameter-setting orders using the procedure in (13).

(13) Procedure for discovering viable parameter-setting orders

(a) Calculate probabilities of encountering unambiguous data for each parameter value, given the current knowledge of the metrical phonology system. This step will produce probabilities like those in tables 2 and 3.

(b) Choose one parameter to set. The value chosen will be the one with the higher probability in the data set, since this is the one a probabilistic learner will eventually converge on over time (e.g. QS over QI in Table 2).

(c) Repeat (a)-(b) until all parameters are set.

(d) If the final parameter values chosen are all the English values, this is a viable parameter-setting order.

(e) Repeat for all possible parameter-setting orders.

6. Results: An unambiguously good strategy with some frequency

It turns out that there are in fact some viable parameter-setting orders that will lead a child using unambiguous data to the English grammar if data tokens are used as input, i.e. hearing the same word again counts as another data point (Table 4). Though there are some orders that do not work (Table 5), learning from unambiguous data in English child-directed speech can still lead a child to English. More specifically, a child using cues has 500 viable orders while a child using parsing has 66 viable orders that will yield acquisition success. Given the complex parametric system and the ambiguous, noisy data set, this is no small feat.

[Put Table 4 approximately here: *Examples of viable parameter-setting orders*.]

[Put Table 5 approximately here: *Examples of non-viable parameter-setting orders.*]

However, acquisition does not succeed if data types are used as input, i.e. the frequency of the word does not matter (as has been suggested for some acquisition tasks (Bybee 1995, Bybee & Hopper 2001)). In this case, there are no viable orders because unambiguous data do not exist in the correct relevant quantities at any point during acquisition. The problem turns out to lie with specific sub-parameters, depending on which identification method is used. If cues are used, the probabilities favor QS-VC-L once QS is set and no subsequent knowledge of the English values causes the probabilities to favor QS-VC-H. If parsing is used, the probabilities favor B-Mor once the system is known to be B, and no subsequent knowledge of the English values causes

the distribution to favor B-Syl. All other parameters have unambiguous data probabilities favoring the English values at some point during acquisition.

So what does this mean for the acquirability of this parametric system for English? If children learn from data tokens, then the system is acquirable from unambiguous data. More specifically, if a child using the unambiguous data learning bias sets the parameters in one of the viable orders, that child will converge on English. For instance, taking the first viable order for a cues child in table 4, the child first chooses to set the quantity sensitivity parameter. Unambiguous data probablities favor the QS value over the QI value. The child then chooses to set the quantity sensitivity sub-parameter, and unambiguous data probabilities favor the QS-VC-H value over the QS-VC-L. This process continues until all parameters are set. Setting the parameters in some order that is not viable (such as those in table 5) will lead to the wrong values for English.

If, instead, children heed only the different data types in the input, this suggests English children will not be able to acquire the English grammar from the English data. A few explanations are possible. First, if children do disregard frequency, this parametric system is not acquirable and so not a viable knowledge representation. A contrasting idea is that children are in fact sensitive to frequency, so the problem that occurs when learning from data types is not relevant. A third, more nuanced idea, is that perhaps the full system is not acquirable until children have knowledge of the interaction with the morphology system, if they disregard frequency. Kehoe (1998) finds that English children up to 28 months may still be deciding on values such as QS-VC-H vs. QS-VC-L, and it is likely children of this age do have some knowledge of morphology. Perhaps

knowledge of the morphological system allows them to make finer distinctions in the input, and so to allow for exceptions. This may be useful as some currently problematic words that are only analyzable as the non-English value (e.g. B-Mor) may be recognized as compound words, e.g. *snowman*, which have their own particular rules of stress.

6.1. Representing the knowledge of viable parameter-setting orders

The viable parameter-setting orders represent the knowledge an English child needs for acquisition success, if the child learns from data tokens. However, it is unlikely English children have a listing of viable parameter-setting orders (either 500 or 66) innately available in their minds, and simply choose one at random to learn English. It turns out, quite fortunately, that the viable orders for both the cues and parsing methods can be captured by very small sets of order constraints (Table 6).

[Put Table 6 approximately here: *English order constraints for viable parameter-setting orders.*]

For cues, there are three constraints such that one parameter value must be set before some other parameter value. For instance, the first constraint in Table 6 states that the child must determine that VC syllables are treated as Heavy (QS-VC-H) before determining that the rightmost syllable is extrametrical (Em-Right). The second constraint states that the child must determine that the rightmost syllable is extrametrical (Em-Right) before determining that a metrical foot's size is determined by the number of

syllables it contains (B-Syl). The last constraint states that the child must determine that metrical feet are two units in size (B-2) before determining that a metrical foot's size is determined by the number of syllables it contains (B-Syl).

For parsing, there are three groups such that the first one must be set before the second one, and the second one must be set before the third one. Looking again to Table 6, the child must determine that the language is quantity sensitive (QS), that metrical feet are of some arbitrary bounded size (B), and that metrical feet are headed on the left (Ft-Hd-Left) before determining any of the other parameters of the English grammar. Then, the child must determine that metrical feet are constructed starting from the right edge of the word (Ft-Dir-Rt) and VC syllables are treated as Heavy (QS-VC-H). Finally, the child can determine that the rightmost syllable is extrametrical (Em-Some, Em-Right) and metrical feet are two syllables in size (B-2, B-Syl).

6.2. Comparing the parameter-setting order constraints

Note that the order constraints presented in Table 6 are derived from successful acquisition of the English data itself, rather than from logical consideration of the complexity of the signals for unambiguous data, which is the approach taken in Dresher (1999)'s cue learner. There is some overlap with Dresher's order constraints, however, depending on the method used to identify unambiguous data. For instance, the order constraints the cues learner here must follow are a subset of the constraints proposed by Dresher (1999) for all languages. The parsing learner considered here, on the other hand, requires some order constraints incompatible with the strict ordering Dresher advocates.

For example, Em-Some must be set before Ft-Dir-Rt for Dresher, while Ft-Dir-Rt must be set before Em-Some for the find-all-parses learner here.

The order constraints presented here are meant to apply to children learning English from English child-directed speech, and may not necessarily apply to children learning other languages. Indeed, it is an empirical question for each language whether the acquisition of the parametric system is possible, what order constraints are required if so, and if those order constraints are the same ones as were found for English. There are ways in which to derive order constraints from properties of the acquisition system and the child's previous experience with the language, which will be discussed in section 7.2. The main point is that if languages differ on the necessary order constraints, it may be that the differing order constraints can be derived from some other source that is also variable across languages. In contrast, ordering constraints that are constant across languages (and not derivable through other means) may represent true innate biases children must have in order to acquire the correct parametric system of their language. In general, however they may come to be known by the child, order constraints represent knowledge the child needs to succeed at acquisition for this parametric system.

6.3. Results summary

The main result we find is that the English parametric metrical phonology system is acquirable using data perceived as unambiguous, provided children are sensitive to data frequency and are constrained in the orders in which they set their parameters. This result is pleasantly surprising given the complexity of the system, as well as the

ambiguity and the noisiness of the data. For English, children can identify unambiguous data points, and importantly, identify them in the correct relative quantities. This supports the viability of using an unambiguous data bias for acquisition of parametric systems, since a child selectively learning in this manner can in fact reach the target grammar. It also supports the viability of the parametric system under these acquisition conditions.

7. Discussion

7.1 Order constraints: Why?

In the results, we saw that a child who has a selective learning bias additionally requires prior knowledge about what order to set parameters in. One might wonder why this should be. In the discussion of the data in section 3, we reviewed its noisy nature – specifically, that 27% of the data tokens are incompatible with the target grammar. Still, this leaves 73% that are in fact compatible.

However, the data difficulties are more insidious for an unambiguous data learner. Just because 73% are compatible does not mean that 73% are *unambiguous*. In fact, we saw in section 5 that only a small percentage of the available data are unambiguous for a given parameter value (and of course, none are likely to be unambiguous for all parameter values simultaneously (Clark 1994, among others)). The data unambiguous for a given parameter value depend on the knowledge the child has about the rest of the grammar. It may even be the case that at a certain point in the learning trajectory, no data are unambiguous for the correct value. Thus, what order constraints do is force children into knowledge states where they perceive unambiguous data in favorable probabilities

for acquiring the target grammar. If the child sets a parameter out of order, then the unambiguous English data probabilities will favor the incorrect parameter value. This is because the child's knowledge at that time will cause the child to identify unambiguous data more often for the incorrect parameter value than for the correct one.

Of course, since this seems to be a problem endemic to an unambiguous data learner, we might ask if order constraints could be discarded if the child learned from all the available data instead. Interestingly, results with simulations of unconstrained probabilistic learners (Pearl, to appear) show decided failure to converge on English given English child-directed speech, no matter what order parameters are set in.

Two factors likely contribute to the failure of unbiased learners. First, unbiased learners are implicitly driven by unambiguous data. In probabilistic models in general, unambiguous data will have more impact on the child's beliefs because such data are more informative, by definition. Even if ambiguous data have some influence, unambiguous data will have more influence. So, while unbiased probabilistic learners are not preferentially learning from unambiguous data, they are still be strongly influenced by the unambiguous data in the input. This means that noise that occurs in the unambiguous data afflicts unbiased learners as well.

Second, it turns out that the English child-directed speech data are actually slightly more compatible with other grammars in the hypothesis space. The more compatible grammars are on average compatible with 1.5% more data types and 0.5% more tokens than the English grammar, with the best grammar compatible with 8.2% more types and 3.5% more tokens than the English grammar. This suggests that an unbiased probabilistic

learner – *any* unbiased learner, no matter how sophisticated – is naturally led to those other grammars. So, some kind of bias is needed to drive a probabilistic learner to the English grammar for this parametric system, given English child-directed speech. The unambiguous data bias presented here is one such bias.

7.2 Order constraints as derived knowledge

A child learning from unambiguous data requires parameter-setting order constraints to converge on the English grammar. This requires a certain amount of explicit prior knowledge – specifically, what those order constraints are. It turns out that some order constraints may be derivable from more general properties of the acquisition system, so that the child would naturally follow these order constraints without needing to know them outright. The three properties we will consider will be data saliency, data quantity, and default values. There may in fact be more, but these three come to mind as being fairly general properties of the acquisition system.

Data saliency refers to the inherent "noticeability" of the information – data that are better signals might be noticed more easily by the child. For metrical phonology, the presence of stress may be more salient than absence of stress for simple acoustic reasons: a stressed syllable is more prominent than an unstressed syllable, and so might be more readily attended to. Evidence from morphological rules also suggests that presence of stress is psychologically more salient: there are rules restricting affix attachment to words with stress on a specific syllable (e.g. *–al* for final stress words: *remove* + *al* = *removal*), but there do not seem to be corresponding rules for words without stress on the

appropriate syllable (Bill Idsardi, pers.comm.). With respect to the metrical phonology parameters considered in this study, data saliency may lead to later learning of parameters that require the child to notice the absence of stress on particular syllables (specifically, the extrametricality parameters: Em-Some, Em-Right, Em-Left).

Data quantity refers simply to the amount of perceived unambiguous data available in the input. Parameters with more unambiguous data available are likely to be set before parameters with less, simply because there is more data for them and so the probability of encountering unambiguous data for these parameters is higher. This is very naturally cashed out in any probabilistic model.

Default values are initial assumptions the child will make about a parameter's value (e.g. bounded metrical foot size is determined by syllables (B-Syl), rather than by moras (B-Mor)). These could result from considering which is the simpler hypothesis for a parameter, e.g. assuming B-Syl over B-Mor since the word is already being analyzed by syllables. Order constraints involving parameters with default values may disappear (depending on the order constraint) if the default value is the correct value for the target language (see 7.3.1 for specific examples of this). In addition, the child may bring biases for parameter values from prior experience with the language. While obviously these biases are acquired from the language itself, they function as prior knowledge by the time the child would be acquiring the parametric system discussed here. Note that under this view, metrical phonology acquisition is partitioned into two stages. The first stage includes acquisition of rhythmic properties of the language, but not of the parametric system. In the second stage, the parametric system is acquired, and the child may draw on

knowledge gained during the first stage. For metrical phonology, infant research has shown that children know some of the rhythmic properties of their language even before word segmentation is reliable. Jusczyk *et al.* (1993) demonstrate that English infants have a preference for strong-weak syllable clusters (Ft-Hd-Left) over weak-strong syllable clusters (Ft-Hd-Rt). Kehoe (1998) also suggests 22-month-olds know this about English. Turk *et al.* (1995) show that English infants are sensitive to syllable weight for stress contours (QS). Kehoe (1998) also suggests 22-month-olds seem to have this knowledge about English. So, due to an English child's prior experience with English, we might expect Ft-Hd-Left and QS to be set earlier than other parameters.

7.3 Deriving constraints for cues and parsing

7.3.1 A cues learner

Recall from table 6 that an unambiguous learner using cues must follow three order constraints. First, the child must discover that VC syllables are Heavy (QS-VC-H) before discovering the rightmost syllable is extrametrical (Em-Right). This could fall out from data saliency, since Em-Right requires the child to notice the absence of stress on the rightmost syllable. In contrast, QS-VC-H requires the child to notice the presence of stress in a particular pattern (see table 1 for the QS-VC-H cue pattern).

Second, the child must discover the rightmost syllable is extrametrical (Em-Right) before discovering metrical feet are counted by syllables (B-Syl). This is because the B-Mor value is favored by the unambiguous data probabilities until Em-Right is set. If we look to data quantity, Em-Right has over 20 times as much data as B-Syl, so the child is

20 times as likely to encounter Em-Right data. In this way, the child could learn Em-Right before B-Syl. As another alternative, B-Syl could be viewed as the default value since words are already delimited by syllables (rather than moras). So, if the child initially assumed B-Syl for English, this would be correct and it would not matter when Em-Right was set.

Third, the child must discover that metrical feet are two units long (B-2) before discovering metrical feet are counted by syllables (B-Syl). This is again because B-Mor is favored by the unambiguous data probabilities until B-2 is set. If we look to data quantity, a partial ordering is available that can help. Once Em-Right is set, unambiguous B-2 data is 4 times as probable as unambiguous B-Syl data. Em-Right is over 270 times as probable as B-2 and B-Syl, so data quantity could lead it to be set first. Then, B-2 would be set, followed by B-Syl. An alternative is again having B-Syl as the default hypothesis, so that it is initially set correctly.

Strikingly, we see that all the order constraints a cue learner requires can potentially be derived from properties of the acquisition system, so they do not need to be explicitly known by the child beforehand. The English child using cues to identify unambiguous data can use data saliency, data quantity, and default values to follow a viable parameter-setting order that will lead to the English grammar.

7.3.2 A parsing learner

Recall from table 6 that there are three groups that are ordered with respect to each other. The parameters in the first group (QS, Ft-Hd-Left, B) must be set before those in

the second group (Ft-Dir-Rt, QS-VC-H), which must in turn be set before those in the third group (Em-Some, Em-Right, B-2, B-Syl).

Data saliency can account for the presence of the extrametricality parameters in the last group (Em-Some, Em-Right) – they require the child to notice the absence of stress, and so have less salient data. Thus, they could feasibly be learned later than other parameter values. Prior experience with the language may give the child a head start on Ft-Hd-Left and QS for reasons discussed in section 7.2, and so these parameters' appearance in the first group could also be plausibly derived. Default values, unfortunately, cannot be used with a parsing learner for reasons discussed already (see section 4.3). However, even supposing that they could, only B-Syl seems to be a natural initial hypothesis. It is not clear there is a simple principled explanation for why a learner might initially assume metrical feet are of a specific arbitrary size (Bounded), VC syllables are Heavy in a quantity sensitive system (QS-VC-H), metrical foot construction begins from the right edge of the word (Ft-Dir-Rt), or metrical feet are two units long (B-2). Unfortunately, data quantity will not separate the remaining parameters into the necessary groups (B before QS-VC-H and Ft-Dir-Rt; all of those before B-2). A parsing learner would need this partial ordering explicitly known beforehand.

So, in order to follow a viable parameter-setting order, a find-all-parses child can rely in part on data saliency, data quantity, and prior knowledge of the rhythmic properties of English. However, some specific order information is still required to be known beforehand in order for the child to succeed at acquiring English.

7.3.3 Cues vs. parsing for identifying unambiguous data

Comparing the two methods of identifying unambiguous data, we once again see a complementary set of requirements. A cues learner must know what the cues are in order to find unambiguous data, but does not need any explicitly listed order constraints to converge on English. A find-all-parses learner does not need prior knowledge to identify unambiguous data, but requires certain partial parameter-setting orders to be known beforehand. Thus, there are both benefits and drawbacks to each of these implementations of the unambiguous data learning bias.

It would of course be very useful to combine cues and parsing to capitalize on their complementary strengths and mitigate their complementary weaknesses. One way might be to have the child conduct a limited parse over portions of words, rather than entire words. This is similar in spirit to the partial parsing seen in Fodor & Sakas (2004)'s Waiting Structural Triggers Learner – there, it is ambiguity that triggers a partial parse (of the initial portion of the structure, in those examples). For the metrical phonology case here, the partial parse could be undertaken simply because only a portion of the data point is available (perhaps due to time, attention, or other mental resources available on the part of the child). Since cues are usually pieces of highly informative surface structure that are smaller than the entire word, a child doing limited parsing may be able to derive cue-like structures. Thus, the cues would not need to be specified beforehand and the limited parsing may be less resource-intensive than the full parsing examined in this study. As an explicit example, consider a child encountering an unstressed VV

syllable, and positing that this syllable is at the end of the word (perhaps based on the prosody of the speaker, or transitional probability of the syllables, or both).

(14) Data point encountered (# = word boundary): …VV #

Suppose the child also realizes that the system is quantity sensitive (QS), so syllables classified as Heavy should receive stress. Since a VV syllable will be classified as Heavy, the lack of stress on this syllable will only be acceptable if the Em-Right value is used to parse this data point. So, this data point would signal the necessity of having the rightmost syllable as extrametrical. Thus, by conducting a limited parse over a subpart of this word, the child can derive the cue for Em-Right – a stressless Heavy syllable at the right edge of the word. Many of the cues in Table 1 can be derived in a similar manner.

Assuming that a child using cues derived from limited parsing could succeed at acquiring English, this learning method would potentially have the desired combined strength: (1) less necessarily prior knowledge since the cues are derived, and (2) less resource-intensive identification of unambiguous data, since the parse is over only a subpart of the data point. This prediction remains to be explored.

7.4 Predictions for the acquisition trajectory

If children are learning this parametric system from unambiguous data (identified by either cues or parsing), we would expect them to follow the parameter-setting order constraints laid out in table 6. For instance, whether the child is using cues or parsing, we

would predict quantity sensitivity (QS) to be known before extrametricality (Em-Some, Em-Right). Though it is not currently clear when the knowledge of extrametricality is acquired for English-speaking children, previous research on infants (Turk *et al.* 1995) and on young children (Kehoe 1998) suggests that the bias for quantity sensitivity may already be in place quite early. In addition, we would predict errors persisting longer for the parameters that are set later, e.g. children may make errors on extrametricality even while they demonstrate knowledge of quantity sensitivity.

8. Conclusion

The results obtained from the case study here suggest that a selective learning strategy, in the form of an unambiguous data bias, causes a parametric system of metrical phonology to be acquirable. This provides support for this system as a representation of children's knowledge. One crucial aspect of such a bias is that data are unambiguous relative to the child's perspective, and the child has incomplete knowledge of the full adult grammar during the acquisition process. Thus, the informativity of the data point will change depending on the child's current knowledge of the system. For English metrical phonology, this seems to provide the flexibility a child needs to succeed. However, this success does rest on the child being sensitive to the frequency of the data types encountered. If children learn only from data types and not from tokens, then the full parametric system is not in fact acquirable. Still, if children are indeed sensitive to frequency, we have also generated predictions for their acquisition trajectory that can be verified with experimental studies. Open research questions do of course remain for

several of the ideas considered here: (a) the success of the unambiguous data bias for other languages and other parametric systems, (b) other methods of implementing a selective learning bias (e.g. see Yang (2005) for learning only from systematic data) and the success of such biases, (c) a detailed account of the acquisition of both the core parametric system and the irregular data patterns (see Yang (2002) for acquisition of English past tense morphology, which has similar irregularities), and (d) the acquirability of other knowledge representations, such as constraint-ranking systems (Tesar & Smolensky 2000) and FSA-style grammars (Heinz 2007). The domain of metrical phonology seems a very fruitful domain in which to answer these questions.

Acknowledgements

I am very grateful to Amy Weinberg, Bill Idsardi, Jeff Lidz, Charles Yang, Walter Daelemans, five anonymous reviewers, and the audience at BUCLD 2007 for encouragement and thoroughly sensible suggestions. If I failed to follow any of their advice here, it's entirely my own fault. This research was supported by NSF grant BCS-0843896.

Appendix. Child-directed speech data and adult-directed speech data

The child-directed speech data comprising the corpus used as 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. Stress patterns absent from the table have a token and type frequency of 0 in this corpus.

[put Table A1 approximately here: *Child-directed speech data.*]

It is reasonable to ask if the child-directed speech data differ with respect to stress contour from adult-directed speech data. Figures A1 and A2 show comparisons by tokens and types respectively for the child-directed speech corpus above to the North American English portion of the CALLFRIEND corpus (Canavan & Zipperlen 1996) available from TalkBank (http://talkbank.org/). The CALLFRIEND corpus contains transcripts of phone calls between English speakers in the United States, and the North American English portion consists of 82485 tokens and 4720 types.

We can first notice that there seems to be reasonable overlap in terms of how many words of *n* syllables comprise the corpus, if we count tokens. In addition, the 2-syllable words tend to have similar distributions in terms of what stress contours are represented

in them. However, the other word distributions are significantly different, especially when comparing data types. This suggests the use of specifically child-directed speech is important for testing acquirability.

[put Figure A1 approximately here*: Distribution comparison by data tokens.*]

[put Figure A2 approximately here: *Distribution comparison by data types.*]

Endnotes

[1] Again, note that other implementations of parsing (Sakas 2003, Fodor & Sakas 2004) may not have the strengths and weaknesses detailed here for find-all-parses parsing. Instead, they resemble cues more in their ability to identify useful data easily and learn from subparts of a data point. However, these implementations also resemble cues in their inability to identify data as unambiguous with absolute certainty. Often, such methods will learn from ambiguous data, guessing when there is uncertainty.

[2] See Sakas & Fodor (2001), who acknowledge these problems and propose ways to solve them in scenarios where the adult language data does not contain non-trivial quantities of irregular data.

[3] However, it may be possible to sidestep this problem with a different instantiation of parsing, specifically a probabilistic parser that favors default values and probabilistically uses them for parsing (see Yang (2002) and Fodor & Sakas (2004) for examples of probabilistic parsing for learning). The learner would still be able to occasionally parse the unambiguous data encountered for the non-default value – it would simply use this

value with low probability. Note that this instantiation will not necessarily find truly unambiguous data since not all parameter value combinations are tried on each data point.

References

Archibald, J. (1992). Adult abilities in L2 speech: evidence from stress. (In J. Leather & A. James, (eds.) New Sounds 92: Proceedings of the 1992 Amsterdam Symposium on the Acquisition of Second Language Speech, (pp.1-16). Amsterdam: University of Amsterdam Press)

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

Brent, M. and Siskind, J. (2001). The Role of Exposure to Isolated Words in Early Vocabulary Development. Cognition, 81/82:33–44.

Bybee, J. (1995). Regular Morphology and the Lexicon. Language and Cognitive Processes, 10(5), 425-455.

Bybee, J. & Hopper, P. (2001). Frequency and the Emergence of Language Structure. (Amsterdam: John Benjamins).

Canavan, A., Graff, D., and Zipperlen, G. (1997). CALLHOME American English Speech. (Philadelphia, PA: Linguistic Data Consortium)

Canavan, A., and Zipperlen, G. (1996). CALLFRIEND American English-Non-Southern Dialect. (Philadelphia, PA: Linguistic Data Consortium)

Chomsky, N. (1981). Lectures on Government and Binding. (Dordrecht: Foris)

Chomsky, N. & Halle, M. (1968). The sound pattern of English. (New York: Harper and Row)

Clark, R. (1994). Kolmogorov complexity and the information content of parameters. IRCS Report 94-17. Institute for Research in Cognitive Science, University of Pennsylvania.

Daelemans, W., Gillis, S., and Durieux, G. (1994). The Acquisition of Stress: A Data-Oriented Approach. Association for Computational Linguistics, 20(3), 421-451.

Dresher, B. E. (1999). Charting the Learning Path: Cues to Parameter Setting. Linguistic Inquiry, 30: 27-67.

Dresher, B. E. and Kaye, J. (1990). A computational learning model for metrical phonology. Cognition, 34:137-195.

Fodor, J. D. (1998a). Unambiguous Triggers. Linguistic Inquiry, 29: 1-36.

Fodor, J. D. (1998b). Parsing to Learn. Journal of Psycholinguistic Research, 27(3): 339-374.

Fodor, J.D. & Sakas, W. (2004). Evaluating models of parameter setting. (In A. Brugos, L. Micciulla and C. E. Smith (eds.), BUCLD 28: Proceedings of the 28th Annual Boston University Conference on Language Development. Somerville, MA: Cascadilla Press)

Foraker, S., Regier, T., Khetarpal, A., Perfors, A., and Tenenbaum, J. (2007). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. (In D. S. McNamara and J. G. Trafton (eds.), Proceedings of the 29th Annual Conference of the Cognitive Science Society)

Gibson, E. and K. Wexler (1994). Triggers. Linguistic Inquiry, 25: 355-407.

Goldberg, A. (1995). Constructions: A construction grammar approach to argument structure. (Chicago, IL: University of Chicago Press)

Goldwater, S., Griffiths, T., and Johnson, M. (2007). Distributional Cues to Word Segmentation: Context is Important. (In Caunt-Nulton, H., Kulatilake, S., and Woo, I. (eds.), BUCLD 31: Proceedings of the 31st Boston University Conference on Language Development)

Halle, M. & Idsardi, W. (1995). General Properties of Stress and Metrical Structure. (In Goldsmith, J. (ed.), The Handbook of Phonological Theory, (pp. 403-443), Cambridge, MA & Oxford: Blackwell Publishers)

Halle, M. & Verganud, J-R. (1987). An Essay on Stress. (Cambridge, MA: MIT Press)

Hayes, B. (1980). A Metrical Theory of Stress Rules. Dissertation, M.I.T.

Hayes, B. (1982). Extrametricality and English stress, Linguistic Inquiry, 13, 227-276.

Hayes, B. (1995). Metrical Stress Theory: Principles and Case Studies. (Chicago: University of Chicago Press)

Hayes, B. and Wilson, C. (2008). A Maximum Entropy Model of Phonotactics and Phonotactic Learning. Linguistic Inquiry, 39, 379-440.

Heinz, J. (2007). Learning Unbounded Stress Patterns via Local Inference. (In Proceedings of the 37th Annual Meeting of the Northeast Linguistics Society (NELS 37))

Hochberg, J. (1988). Learning Spanish Stress: Developmental and Theoretical Perspectives. Language, 64(4), 683-706.

Johnson, K. (2004). Gold's Theorem and Cognitive Science. Philosophy of Science, 71, 571-592.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress patterns of English words. Child Development, 64, 675-687.

Kehoe, M. (1998). Support for metrical stress theory in stress acquisition. Clinical Linguistics & Phonetics, 12(1), 1-23.

Kiparsky, P. (1979). Metrical Structure Assignment is Cyclic. Linguistic Inquiry, 10.4, 421-441.

Lightfoot, D. (1999). The development of language: Acquisition, change and evolution. (Oxford: Blackwell).

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

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

Pearl, L. (2007). Necessary Bias in Natural Language Learning. Dissertation, University of Maryland.

Pearl, L. (to appear) Learning English Metrical Phonology: Beyond Simple Probability. (Proceedings of Generative Approaches to Language Acquisition North America 3, University of Connecticut.)

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

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the Stimulus? A rational

approach. (In Proceedings of the 28th Annual Conference of the Cognitive Science Society)

Prince, A. and Smolensky, P. (2004). Optimality Theory: Constraint Interaction in Generative Grammar. (Blackwell)

Sakas, W. (2000). Ambiguity and the Computational Feasibility of Syntax Acquisition. Dissertation, City University of New York.

Sakas, W., and Fodor, J.D. (2001). The Structural Triggers Learner, in S. Bertolo (ed.), Language Acquisition and Learnability, 172-233, Cambridge University Press.

Sakas, W. & Nishimoto, E. (2002). Search, Structure, or Statistics? A Comparative Study of Memoryless Heuristics for Syntax Acquisition. Ms., CUNY: New York.

Sakas, W. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. (In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: ACL)

Shannon, C. (1948). A mathematical theory of communication, Bell System Technical Journal, 27, 379-423 and 623-656.

Tenenbaum, J. & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. Behavioral and Brain Sciences, 24, 629-640.

Tesar, B. & Smolensky, P. (2000). Learnability in Optimality Theory. (Cambridge, Massachusetts: The MIT Press)

Tomasello, M. (2006). Acquiring linguistic constructions. (In D. Kuhn & R. Siegler (Eds.), Handbook of Child Psychology. New York: Wiley)

Turk, A., Jusczyk, P. & Gerken, L. (1995). Do English-learning Infants Use Syllable Weight to Determine Stress? Language and Speech, 38(2): 143-158.

Valian, V. (1990). Null subjects: A problem for parameter setting models of language acquisition. Cognition, 35, 105-122.

Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. (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, M. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. Behavioural Research Methods, Instruments and Computers, 20(1): 6-11.

Yang, C. (2002). Knowledge and Learning in Natural Language. (Oxford: Oxford University Press)

Yang, C. (2004). Universal Grammar, statistics or both? Trends in Cognitive Science, 8(10), 451-456.

Yang, C. (2005). On productivity. Yearbook of Language Variation, 5, 333-370.

Table 1. Cues for metrical phonology parameter values. Some cues may depend on the child's current knowledge state, represented in *italics*. For example, the cue for QS depends on what is known about extrametricality (*Em-None/Em-Some/Em unknown*).

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

	stressed L syllable at right edge (…H L)
	QS, Em-Right: stressed H syllable followed by stressed L syllable followed by unstressed syllable at right edge (…H L H)
Ft-Dir-Rt	*QI or Q-unknown, Em-None/Right or Em unknown*: 2 stressed adjacent syllables at left edge (VC V…)
	QI or Q-unknown, Em-Left: unstressed syllable followed by 2 stressed adjacent syllables at left edge (VC V VV…)
	QS, Em-None/Right or Em unknown: stressed L syllable followed by stressed H syllable at eft edge (L H…)
	QS, Em-Left: unstressed syllable followed by stressed L syllable followed by stressed H at left edge (H L H…)
Unb	*QI or Q-unknown*: 3+ unstressed syllables in a row (…VC VV VC…)
	QS: 3+ unstressed Light syllables in a row (…L L L)
B	Union of B-2 and B-3 cues
B-2	*QI or Q-unknown*: 3+ syllables in a row, every other one stressed (… VC VV VC…)
	QS: 3+ Light syllables in a row, every other one stressed (…L L L…)
B-3	*QI or Q-unknown*: 4+ syllables in a row, every third one stressed (…V VC VV V…)
	QS: 4+ Light syllables in a row, every third one stressed (…L L L L…)
B-Syl	*QI or Q-unknown*: Union of QI B-2 and QI B-3 cues
	QS, B-2: 2 adjacent syllables, one stressed Heavy and one unstressed

	Light (…H L…)
	QS, B-3: 3 adjacent syllables, 2 unstressed Light preceding a stressed Heavy or following a stressed Heavy (…H L L…), (…L L H…)
B-Mor	*Em-None or Em-unknown*: 2 syllable word with both syllables Heavy and stressed (H H)
	Em-Some: 3 syllable word with 2 adjacent syllables Heavy and stressed (L H H)
Ft-Hd-Left	*Em-None or Em-unknown*: Leftmost syllable is stressed (VC…)
	Em-Left: 2nd from leftmost syllable is stressed (VV VC…)
Ft-Hd-Rt	*Em-None of Em-unknown*: Rightmost syllable is stressed (…VC)
	Em-Right: 2nd from rightmost syllable is stressed (…VC VV)

Table 2. Initial probabilities of unambiguous data. Probabilities are quite small, since much data is ambiguous to the child at this point.

Quantity Sensitivity		**Extrametricality**	
QI	QS	Em-None	Em-Some
0.00398	0.0205	0.0284	0.0000259
Feet Directionality		**Boundedness**	
Ft-Dir-Left	Ft-Dir-Rt	Unb	B
0.000	0.00000925	0.00000370	0.00435
Feet Headedness			
Ft-Hd-Left	Ft-Hd-Rt		
0.00148	0.000		

Table 3. Probabilities of unambiguous data after QS is set. Probabilities for many parameter values have shifted, especially those for the extrametricality parameters. The learner also explores the QS sub-parameter that decides how to treat VC syllables.

QS VC Syllables		**Extrametricality**	
QS-VC-L	QS-VC-H	Em-None	Em-Some
0.00265	0.00309	0.0240	0.0485
Feet Directionality		**Boundedness**	
Ft-Dir-Left	Ft-Dir-Rt	Unb	B
0.000	0.00000555	0.00000370	0.00125
Feet Headedness			
Ft-Hd-Left	Ft-Hd-Rt		
0.000588	0.0000204		

Table 4. Examples of viable parameter-setting orders when learning from data tokens. Orders are read left to right, with the parameter values to the left being set before parameter values to the right.

Cues
(1) QS, QS-VC-H, B, B-2, Ft-Hd-Left, Ft-Dir-Rt, Em-Some, Em-Right, B-Syl
(2) B, B-2, Ft-Hd-Left, Ft-Dir-Rt, QS, QS-VC-H, Em-Some, Em-Right, B-Syl
(3) Ft-Hd-Left, Ft-Dir-Rt, QS, QS-VC-H, B, Em-Some, Em-Rt, B-2, B-Syl
Parsing
(1) B, QS, Ft-Hd-Left, Ft-Dir-Rt , QS-VC-H, B-Syl, Em-Some, Em-Right, B-2
(2) Ft-Hd-Left, QS, QS-VC-H, B, Ft-Dir-Rt,, Em-Some, Em-Right, B-Syl, B-2
(3) QS, B, Ft-Hd-Left, QS-VC-H, Ft-Dir-Rt, B-Syl, Em-Some, Em-Rt, B-2

Table 5. Examples of non-viable parameter-setting orders. Orders are read left to right, with the parameter values to the left being set before parameter values to the right. Incorrect parameter values are in ***bold italics***. All orders continuing on from these incorrect values will converge on incorrect grammars.

Cues
(1) QS, QS-VC-H, B, B-2, ***B-Mor***, …
(2) B, B-2, Ft-Hd-Left, ***B-Mor***, …
(3) ***Em-None***, …
(4) Ft-Hd-Left, ***Em-None***, …
Parsing
(1) QS, QS-VC-H, B, B-Syl, B-2, Em-Some, Em-Right, ***Ft-Hd-Rt***…
(2) B, B-Syl, B-2, ***Em-None***, …
(3) ***Em-None***, …
(4) Ft-Hd-Left, ***Ft-Dir-Left***, …

Table 6. English order constraints for viable parameter-setting orders.

Cues: Follow these constraints, other parameters freely ordered
(1) QS-VC-H before Em-Right
(2) Em-Right before B-Syl
(3) B-2 before B-Syl
Parsing: Group 1 before Group 2, Group 2 before Group 3
Group 1: QS, B, Ft-Hd-Left
Group 2: Ft-Dir-Rt, QS-VC-H
Group 3: Em-Some, Em-Right, B-2, B-Syl

Table A1. Child-directed speech data.

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) 0110: (2 / 1) 1100: (20 / 8) 0101: (18 / 14) 1010: (910 / 161) 0100: (50 /26) 1001: (7 / 3)	Stress Pattern Frequency 11010: (1 / 1) 01100: (1 / 1) 10101: (54 / 1) 01010: (67 / 8) 10100: (39 / 15) 01000: (2 / 1)
6-syl: (4 / 4)	7-syl: (1 / 1)
Stress Pattern Frequency 100100: (2 / 2) 010010: (1 / 1) 100010: (1 / 1)	Stress Pattern Frequency 1010100: (1 / 1)

Figure A1. Distribution comparison by data tokens.

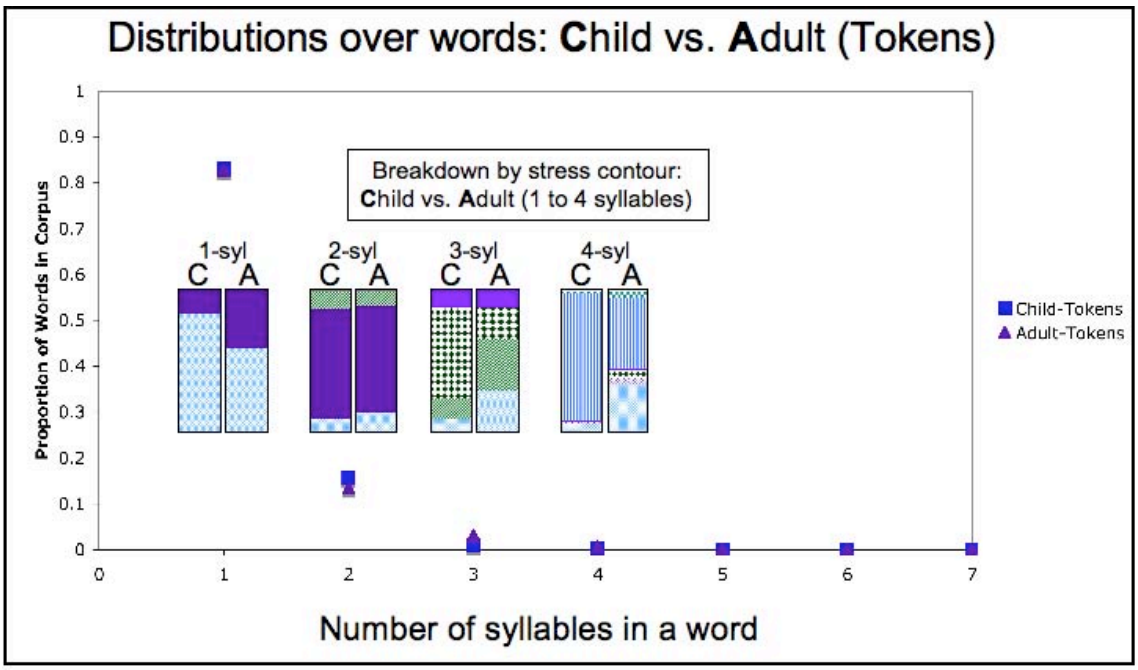

Figure A2. Distribution comparison by data types.