Parameters in Language Acquisition

Lisa Pearl and Jeffrey Lidz

1. What a parameter is (or is meant to be), and what it's for

A parameter, in it simplest conception, is an abstraction that can account for multiple observations in some domain. A parameter in statistical modeling, for example, determines what the model predicts will be observed in the world in a variety of situations; a parameter in our mental model would determine what *we* predict will be observed in the world in different scenarios.

In statistical modeling, there are numerous parameters defined mathematically to account for the expected shape of data distributions. Take the Gaussian distribution, sometimes called the "normal distribution" or "bell curve", which can be used as a simple statistical model to explain complex phenomena (such as "how many minutes late to class I'll be"). In this model, the value of the measured variable (X in figure 1 below) tends to cluster around the mean μ. Two parameters are used to determine the shape of the curve that represents the expected distribution of data: the mean μ and the variance σ^2. These parameters are part of the fixed function φ(X) (note the y axis label in figure 1) that produces the probability of X having a specific value:

$$(1)\ \varphi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

While the function itself is invariant, the predictions produced with this function vary based on the values of μ and σ^2. Changing the value of these parameters demonstrably changes the expectation of how many data points with particular values we expect to see (or comparably, how often we expect to see a data point with a particular value).

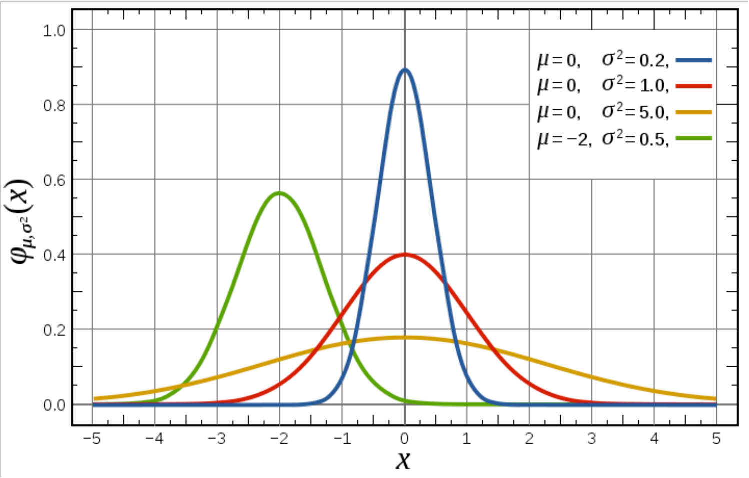

Figure 1. The Gaussian distribution, with different values of μ and σ^2. Courtesy of Wikipedia Commons.

In a related fashion, imagine that we are trying to determine the values of μ and σ^2 in the function φ(X) for some data that we observe. Observing different quantities of data with particular values can tell us which values of μ and σ^2 are most likely. For instance, if we observe 1000 data points whose values are mostly between -3 and -1 (perhaps producing a curve like the green one in figure 1), we would be more likely to think that μ is -2 rather than 0, and that smaller values of σ^2 are more plausible than large values. Even if we have never seen a data point with value 5, we can expect that this data point is not very likely to occur, given what we have inferred about the data distribution and how we can describe that data distribution with the values of μ and σ^2. Importantly, we do not see the process that generates the data, but only the data themselves. This means that in order to form our expectations about X, we are, in effect, reverse engineering the observable data in order to figure out how these data were generated, given what we (perhaps unconsciously) know about function φ(X) and variables μ and σ^2. Our knowledge of the underlying function and the associated parameters that generate these data allows us to represent an infinite number of expectations about the behavior of variable X. Indeed, within the cognitive sciences, learning is generally conceived as this kind of inverse problem: given a set of observations (e.g., sentences), learning consists of identifying the underlying system that determined those observations (e.g., the grammar). (Chomsky 1965, Gallistel 1990, inter alia). Notably, and perhaps particularly relevant for the nativist view of language, knowing the function allows us to generate an infinite number of expectations that are nonetheless constrained in their behavior (here, constrained to obey this underlying function). Thus, the hypothesis space consisting of the infinite expected behaviors, given this function, is still smaller than the hypothesis space consisting of all possible behaviors. Knowing the function thus narrows the hypothesis space of expected behaviors, while still allowing an infinite number of expected behaviors.

Meanwhile, statistical parameters of a function can help us describe complex phenomena in a very compact way. By knowing one function and the values of only two parameters, we can define a Gaussian data distribution that predicts how likely we are to see a data point with a particular value ("I'll be 15 minutes late to class", X=15), without necessarily observing that particular data point a large number of times – or even at all. This, of course, is just one example of parameters used in statistical modeling. There are numerous others, some of which have been recently used in statistical modeling of language acquisition (see, for example, Tenenbaum, Griffiths, and Kemp (2006), Goldwater, Griffiths, and Johnson (2009), and Perfors, Tenenbaum, and Wonnacott (2010) for examples of hierarchical Bayesian modeling that use a small number of parameters).

In the domain of human language, both principles and parameters are often thought of as innate domain–specific abstractions that connect to many structural properties about language. Linguistic principles correspond to the properties that are invariant across all human languages. Using our statistical analogy from before, the equation's form is invariant – it is the statistical "principle" that explains the observed data. Linguistic parameters correspond to the properties that vary across human languages. Again using our statistical analogy, the statistical parameters of μ and σ^2 determine the exact form of the curve that represents the likelihood of observing certain data. While different values for these parameters can produce many different curves, these curves share their underlying form due to the common invariant function in (1).

Since linguistic principles do not vary, they might not need to be learned by the child. Parameter values, on the other hand, must be inferred from the language data since languages do not all have the same parameter values. As with the statistical parameters above, children must reverse engineer the observable data to figure out the most likely way these data were generated. Even given knowledge of linguistic principles and linguistic parameters, children must still decide the most likely value for any parameter.

The fact that parameters connect to multiple structural properties then becomes a very good thing from the perspective of someone trying to acquire language. This is because a child can learn about that parameter's value by observing many different kinds of examples in the language. As Hyams (1987) notes, "the richer the deductive structure associated with a particular parameter, the greater the range of potential 'triggering' data which will be available to the child for the 'fixing' of the particular parameter."

This idea figures prominently in hierarchical Bayesian models (see Kemp, Perfors, and Tenenbaum (2007) for an accessible overview), where a learner can make generalizations about the observable data at different levels of abstraction. To take a non–linguistic example from Kemp et al. (2007), suppose a learner must learn something about bags containing marbles. Suppose this learner observed twenty bags, each of which is filled with 100 black marbles or 100 white marbles. This learner then sees a new bag, and is only shown one marble from it, which happens to be purple. A hierarchical Bayesian learner will likely make the following inference at this point: All the marbles in the new bag are purple. The learner can make this inference because it is able to learn something about properties of bags in general, sometimes called an "overhypothesis" (Kemp et al. 2007), instead of only about bags containing black or white marbles. Specifically here, bags seem to contain only one color of marble. That is, by observing individual bag examples, the hierarchical Bayesian learner has learned something about how to compactly represent the structure of those bags' composition – and it infers this property will apply to bags in general. In this way, the more abstract knowledge of marble bag composition has been triggered by many individual examples of marble bags. More interestingly, this learner was able to have certain expectations about a marble bag containing a novel marble color by learning from bags containing only black or white marbles.

Translating this to a linguistic example, suppose the marble bags are individual sentences. Some sentences contain verbs and objects, and whenever there is an object, suppose it appears after the verb (e.g., *see the penguin,* rather than *the penguin see*). Other sentences contain modal verbs and nonfinite main verbs, and whenever both occur, the modal precedes the main verb (e.g., *could see*, rather than *see could*). A learner could be aware of the shared structure of these sentences – specifically that these observable forms can be characterized as the head of a phrase appearing before its complements (specifically [$_{VP}$ V NP] and [$_{IP}$ Aux VP]) – and could encode this "head–first" knowledge at the level that describes sentences in general, akin to a hierarchical Bayesian learner's overhypothesis. This learner would then infer that sentences in the language generally have head–first structure. If this learner then saw a sentence with a preposition surrounded by NPs (e.g. *penguins on icebergs are adorable*), it would infer that the preposition should precede its object ([$_{NP}$ *penguins* [$_{PP}$ *on icebergs*]] and not [$_{NP}$ [$_{PP}$ *penguins on*] *icebergs*]), even if it had never seen a preposition used before with an object. In this way, the notion of a head–directionality parameter (discussed further in section 3) can be encoded in a hierarchical learner. More broadly, parameters are similar in spirit to the overhypotheses in a hierarchical Bayesian learner: Inferences (e.g., about prepositional phrase internal structure) can be made on the basis of examples that bear on the general property being learned (e.g., head directionality), even if those examples are not examples of the exact inference to be made (e.g., verb phrase examples).

Parameters can be especially useful when a child is trying to learn the things about language structure that are otherwise hard to learn, perhaps because they are very complex properties themselves or because they appear very infrequently in the available data. The way it would work is this: each hard–to–learn property is linked to a parameter that has at least one easy–to–observe property connected to it. Through the easy–to–observe property, children learn about the value of the parameter that is connected to it. Through the parameter, the hard–to–learn property follows, perhaps even without seeing any data that directly instantiate it. Thus, the parameter would allow children to observe data that fix the value of the hard–to–learn property. So, structures that may not have been directly observed can nonetheless be generated by the grammar.

If we think about the hypothesis space of a child trying to acquire language, having parameters can make acquisition a lot easier (as argued by Chomsky (1981b, 1986a)). Instead of needing to identify the language properties individually, the child now has a built–in shortcut since some properties are connected to other properties, and can thus be acquired on the basis of fewer (and perhaps less varied) observations. For example, suppose we have 5 different structural properties in the language, each of which is binary. If each property has to be learned individually, there are 2^5 possible ways these 5 structural properties may be set. So, the child's hypothesis space contains 2^5 hypotheses of how the language could work. To learn each property, the child must see data pertaining to each property. (This many not seem so bad right now, but languages clearly have many more properties than this. See section 3 for some suggested parameters.) Suppose now that there is one binary parameter connected to each of these 5 binary language properties. This means that if that one parameter is set to "yes", all 5 connected properties have their values set definitively one way (either to "yes" or to "no", depending on the property). If the parameter is instead set to "no", the 5 connected properties have their values set definitively the other way. So, the child's hypothesis space now contains only 2 hypotheses of how language could work, one with the parameter set to "yes" and one with the parameter set to "no". Moreover, the child can learn about properties 2, 3, 4, and 5 by observing data only for property 1 and learning what the parameter's value is. This would seem to be much easier than learning each property individually.

So, in sum, the word “parameter” is simply the name given to those abstract features of grammar (a) that govern many different observable structures and (b) that vary from language to language. Its use in both the theory of language acquisition and the theory of grammar typology is to condense the representation of the language, thereby structuring the learning task for the child in such a way as to reduce the range of observations required to construct a grammar. In theory, this works by connecting together observations that might otherwise need to be accounted for independently from each other.

2. How do we tell what counts as a parameter?

Parameters seem useful in the abstract, but how do we define what a parameter is? How do we recognize when we have found one?

Chomsky (1981b) imagines parameters much as we have discussed them above: “…we hope to find that complexes of properties…are reducible to a single parameter, fixed in one way or another”. That is, a parameter should connect to many different properties about a language, and not simply apply only for one property. Safir (1987), in fact, cautions against “describing any sort of language difference in terms of some *ad hoc* parameter” in order to prevent a parametric theory from “licensing mere description”. The point of parameters is to make the description of language structure more compact in some sense. Having a parameter for each point of variation in a language wouldn’t accomplish this very well. Smith and Law (2009) note that parameters that connect to a variety of linguistic properties are often referred to as “macro–parameters” while “micro–parameters” are those that connect to only a few properties. Macro–parameters seem to be more in line with the original spirit of linguistic parameters.

Given that positing an independent parameter for each point of variation would undercut the explanatory value of parameters, it is important to determine just what the right parameters are. Ideally, this will come from the theory of comparative grammar and there are many proposals, which vary in their success at capturing the full range of observed variation (Newmeyer 2005).

One of the earliest parameters proposed concerned the nature of overt subjects. Rizzi (1982) argued that a cluster of properties varied predictably with whether overt subjects were obligatory. For example, languages that do not require overt subjects in simple declaratives allow postverbal subjects, do not exhibit expletive subjects, and do not exhibit *that*–trace effects (see 2a below). Languages that require a subject in all simple declaratives, however, do not allow postverbal subjects, do exhibit expletive subjects, and do exhibit *that*–trace effects (see 2b below).

(2) Italian vs. English

(a) Italian

(i) -overt subject:
verrá
3rd–sg–fut–*come*
“He will come”

(ii) +postverbal subject:
verrá Gianni
3rd–sg–fut–*come* *Gianni*
“Gianni will come”

(iii) -expletive subjects:
piove
“3rd–sg–*rain*”
“It’s raining”

(iv) -*that*–trace effect
Che credi che verrá?
who 2nd–sg–*think* *that* 3rd–sg–fut–*come*
“Who do you think that will come”[1]

(b) English

(i) +overt subject:
He will come.
*Will come.

(ii) -postverbal subject:
Jack will come.
* Will come Jack.

(iii) +expletive subjects:
It’s raining.
* Is raining.

(iv) +*that*–trace effect
Who do you think will come?
* Who do you think that will come?

This proposal illustrated the potential benefits of parameters because it solved what was believed at the time to be a poverty of the stimulus problem for *that*–trace effects. Specifically, the problem was that English–learning children could observe the optionality of complementizers in declaratives and even in object wh–questions (3). They might then reasonably extend this optionality to subject wh–questions (2b iv).

(3) Optionality of complementizers

(a) declaratives: Jack thought (that) Lily would eat lunch.

(b) object wh–questions: What did Jack think (that) Lily would eat?

Such an extension would be problematic because complementizers are disallowed in subject wh–questions in the adult English grammar. However, a child with an over–general grammar (which allowed a complementizer in subject wh–questions) would never receive direct evidence that complementizers in subject wh–questions were not allowed – the few subject wh–questions they encountered (which would not have a complementizer) would be compatible with their incorrect grammar. Thus, to the extent that English speakers exhibit the *that*–trace effect, a poverty of the stimulus problem exists because the input data are compatible with a grammar that exhibits *that*–trace effects and also with a grammar that does not.

However, if the presence of *that*–trace effects followed from the same parameter setting that determined the nature of subjects, then there simply would be no learnability problem. The relevant data for determining the obligatory absence of complementizers in subject wh–questions would come from the requirement of overt subjects in declaratives and the necessity of expletive subjects. Observing those data, children would set the parameter to the correct value, and then predict that complementizers were not allowed in subject wh–questions. There would be no overgeneralization based on the optionality of complementizers in declaratives and object wh–questions, because children would already have knowledge pertaining to complementizers in subject wh–questions due to this parameter's value, and would thereby circumvent the poverty of the stimulus problem.

The generalization that *that*–trace effects are linked to the nature of subjects has been questioned (Gilligan 2000, Newmeyer 2005), but for our purposes the example serves to illustrate the potential explanatory power of parameters. A parameter is best posited when (a) a cluster of seemingly unrelated properties varies in a systematic way across languages and (b) the parameter solves potential poverty of the stimulus problems.

Sometimes, suggested parameters only seem to connect to a single simple structural variation (for example, object before verb or verb before object in main clauses, or whether the main stress of a word is on the left or the right). Hopefully, at least one of the following happens: (1) these simple structural parameters turn out to be connected to other structural properties in the language (for example, perhaps object–verb order is connected to the order of heads and complements of phrases in general) and/or (2) these parameters are still a more compact representation of the generative system used to produce the language data than alternative options (for example, perhaps having stress parameters is simpler than individually memorizing stress contours for every word and using analogy to generate new stress contours). By doing either of these things, such parameters may help aid acquisition.

3. Using parameters to learn

Of course, even if the child's hypothesis space is constrained by parameters, this doesn't mean that the language acquisition problem is solved. Parameters do solve the problem of what hypotheses to consider about how a language works, and if they can be set correctly, they may alleviate certain poverty of the stimulus problems. Still, children must decide among the available alternatives – and this has proved to be a difficult decision process to model. As an example, suppose children's hypothesis space is defined by *n* binary parameters. This leads to a hypothesis space of 2^n possible grammars, one for each setting of the *n* parameters. If *n* is small, say 5, perhaps this doesn't seem so bad: 2^5 is only 32 options, after all. But given the variety of linguistic properties that appear cross–linguistically (see Baker (2001) for an accessible survey of some), *n* is likely to be larger than this. If *n* includes only 25 parameters, we already have a hypothesis space of 33,554,432 (2^{25}) grammars, which seems like a much more difficult acquisition problem indeed. To converge on the right grammar in, say, eight years, a child would have to eliminate just under 11,500 potential grammars every day.

The simple fact is that, while parameters may help constrain the hypothesis space, every time we add a new one the hypothesis space doubles in size (e.g., a parameter space with 5 binary parameters is half the size of a parameter space with 6 binary parameters (2^5 = 32, 2^6 = 64)). Clearly, the more we can reduce the number of parameters, the easier the acquisition problem will be. Still, from the learner's perspective, this is better than an unconstrained hypothesis space, where the number of potential grammars is potentially infinite.

Yet the number of parameters is not the only difficulty. Children must also decide among the grammars defined by these parameters using the available data, and this turns out not to be so easy sometimes. Because the parameters are abstract and govern a wide range of surface phenomena, it is not always obvious which data uniquely determine any particular parameter setting (Clark 1992, Gibson and Wexler 1994, Niyogi and Berwick 1996, Pearl 2008, Pearl 2009, Pearl 2011). Some parameters may *interact*, meaning that the effect of one parameter masks the effect of another parameter for a particular data point.

For example, consider the domain of metrical phonology, which describes the system for generating stress contours for words (that is, what makes ‘EMphasis’ different from ‘emPHAsis’). Some researchers (Dresher 1999, Halle and Vergnaud 1987, Halle and Idsardi 1995, among others) believe that stress contours are generated by grouping syllables together into larger units called metrical feet, and then stressing various syllables based on properties of those metrical feet. Languages vary in the way that they form metrical feet and in the way they assign stress to metrical feet. Two parameters commonly associated with this view are extrametricality and feet headedness, shown in (4).

(4) Some metrical phonology parameters

(a) Extrametricality: whether all syllables are included in metrical feet, or whether some are left out (and are therefore “extra”metrical).

(b) Foot Headedness: whether the rightmost or leftmost syllable of a metrical foot is stressed (and so whether the “head” of the metrical foot is to the right or to the left)

Suppose a child is trying to determine both whether her language allows extrametrical syllables and what syllable within a foot is stressed. Suppose she hears the word “giRAFFE”. She might come up with the following analyses for explaining why that stress contour is observed:

(5) Analyses for “giRAFFE"[2]

(a) +extrametrical (leftmost syllable), metrical feet headed on the left: The syllable “gi” is extrametrical and not included in a metrical foot. There is one metrical foot, and it includes “raffe”. As it is the leftmost syllable in the metrical foot, it is stressed.

Analysis:	gi raffe
(i) +extrametrical (left)	gi (raffe)
(ii) foot–headed–left	gi (RAFFE)

(b) +extrametrical (leftmost syllable), metrical feet headed on the right: The syllable “gi” is extrametrical and not included in a metrical foot. There is one metrical foot, and it includes “raffe”. As it is the rightmost syllable in the metrical foot, it is stressed.

Analysis:	gi raffe
(i) +extrametrical (left)	gi (raffe)
(ii) foot–headed–right	gi (RAFFE)

(c) -extrametrical, metrical feet headed on the right: There is one metrical foot, and it includes “gi” and “raffe”. “raffe” is the rightmost syllable of the metrical foot, so it is stressed.

Analysis: gi raffe
(i) -extrametrical (gi raffe)
(ii) feet–headed–right (gi RAFFE)

Here, extrametricality seems to be masking whether the language has metrical feet headed on the left or on the right, and foot headedness seems to be masking whether the language has extrametricality. If the child knew the language was -extrametrical, the foot headedness value would be known (feet headed on the right (4c)); conversely, if the child knew that language had feet headed on the left, then the extrametricality value would be known (+extrametrical on the leftmost syllable (4a)). However, neither value is known, and so it is difficult to tell what parameter combination generated this data point.

We can find a similar situation in syntax if we look at the parameters some researchers have called verb–second movement and head directionality. First, verb–second movement: Several linguists (Chomsky 1981b, Cook and Newson 1996, Guasti 2002, Sakas 2003, Yang 2004, among others) have described a parameter that leads to the transposition of subjects and auxiliaries in English questions (6) and to the tensed verb always being in second phrasal position in languages like German (7a,b). It should be noted that while English has this verbal movement for questions, it does not require the tensed verb to be in second phrasal position in declarative clauses (8). Thus, this verb movement parameter has different effects on observable word order, depending on clause type (declarative vs. interrogative).

(6) Transposition of the subject and auxiliary verb in English question formation
(a) Underlying sentence: The penguin *must* eat fish.
(b) Yes/No question: *Must* the penguin eat fish?

(7) Second phrasal position of the tensed verb in German
(a) Der Pinguin isst Fisch.
the penguin eats fish
“The penguin eats fish”

(b) Fisch isst der Pinguin.
fish eats the penguin
"The penguin eats fish"

(c) Der Pinguin muss Fisch essen.
The penguin must fish eat
"The penguin must eat fish"

(d) Fisch muss der Pinguin essen.
fish must the penguin eat
"The penguin must eat fish"

(8) Non–second phrasal position of the tensed verb in English
(a) The penguin usually eats fish.
(b) Sometimes the penguin eats fish.

The head directionality parameter (Baker 2001, Cook and Newson 1996), concerns the position of phrasal heads (for example, the verb in a verb phrase) with respect to the phrasal complement (for example, the object of the verb in the verb phrase). The idea is that a language consistently has the heads of its phrases all on the same side of the complements of all its phrases, whether the head is first in the phrase (head–first: English, Edo, Thai, Zapotec) (9a) or the head is last in the phrase (head–last: Lakhota, Japanese, Basque, Amharic) (9b).

(9) Head–first vs. head–last languages
(a) English is head–first

for the penguin: the preposition *for* is before its object *the penguin* in the preposition phrase

hugged the penguin: the verb *hugged* is before its object *the penguin* in the verb phrase

(b) Lakhota is head–last

Jack wowapi k'uhe oyųke ki ohlate iyeye.
Jack letter that bed the under found
"Jack found that letter under the bed"

The verb *iyeye* is after its objects (*wowapi k'uhe* and *oyųke ki ohlate*) and the preposition *ohlate* is after its object (*oyųke ki*).

Suppose a German child hears the sentence in (10) and is trying to determine the value of the verb–second and head–directionality parameters.

(10) Example German sentence

Ich	liebe	Pinguine.
I	*love–1^{ST}–SG*	*penguins*
SUBJECT	VERB	OBJECT

"I love penguins."

Here, the verb's complement is the object (*Pinguine*), and the child might come up with the following analyses to explain the observed word order:

(11) Analyses for *Ich liebe Pinguine*

(a) -verb–second, heads precede their complements:

Underlying order:			Ich	liebe	Pinguine
			Subj	Verb	Obj
Observable order:			Ich	liebe	Pinguine
			Subj	Verb	Obj

(b) +verb–second, heads precede their complements:

Underlying order:			Ich	liebe	Pinguine
			Subj	Verb	Obj
Observable order:	Ich	liebe	t_{Ich}	t_{liebe}	Pinguine
	Subj	Verb	t_{Subj}	t_{Verb}	Obj

(c) +verb–second, heads follow their complements:

Underlying order:			Ich	Pinguine	liebe
			Subj	Obj	Verb
Observable order:	Ich	liebe	t_{Ich}	Pinguine	t_{liebe}
	Subj	Verb	t_{Subj}	Obj	t_{Verb}

We again find that one parameter seems be masking the effects of the other. If the child knew the language was -verb–second, then the value of the head directionality parameter would be clear for this data point (heads precede their complements (11a)); conversely, if the child knew heads followed their complements, then the value of the verb–second parameter would be clear for this data point (the language has verb–second movement (11c)). Since both parameters are undetermined, however, extracting the right values for them from this data point is not straight forward.

As we noted in the introduction, the main trouble is that children don't see the process that produces the data – they only see the data themselves. Children must infer the most likely parameter values from the observable data, and the observable data may be (and often are) ambiguous as to which parameter values were used to generate them. So, even if children know the set of parameters and the potential values these parameters might take, acquisition still has some obstacles left.

This leaves open a number of questions about the parameter–setting process, which we explore in the remainder of this section:

- Are parameter settings represented as deterministic or probabilistic? More specifically, at any one stage, does the child maintain a single hypothesis about which way the parameter is set (like a light switch)? Alternatively, does the child maintain multiple hypotheses along with a confidence value (more like a dimmer switch between extremes)?

- Do children make inferences from a large quantity of data all at once, or do they make inferences incrementally as they encounter the data?

- Do children attempt to learn parameters individually, or are parameters learned only as part of an entire grammar that succeeds or fails at analyzing the observable data? If parameters are learned individually (rather than as an ensemble), what causes a change in a child's hypothesis for a parameter value?

- Which data are used to set parameters?

- Do children need to follow particular parameter–setting orders?

- What happens when children do not encounter sufficient informative data?

3.1 Light switches or light dimmers?

Suppose a child is attempting to learn the value for a particular parameter, such as extrametricality in the metrical phonology system. Two main options exist: the language either has extrametricality or it does not. We could imagine at least two ways children might explore the different hypotheses. Perhaps they choose a parameter value to start off with at random, maintain that value as their working hypothesis, and only discard it for the alternative hypothesis if it fails to analyze the input data successfully (Fodor and Sakas 2004, Gibson and Wexler 1994, Niyogi and Berwick 1996, Sakas 2003, Sakas and Fodor 2001, Sakas and Nishimoto 2002). Parameter–setting behaves as a light switch that can only choose one option at a time. One issue with this approach is that it predicts that the child's learning behavior should exhibit abrupt changes with each change in parameter value – and once the correct parameter value is acquired, the child should use it all the time.[3] We would never expect children to go back to using the incorrect hypothesis if they are only maintaining the single correct hypothesis. This does not seem to map well to what we know of children's behavior – the trajectory of acquisition appears to be more gradual, with children intermixing correct and incorrect linguistic behavior for a period of time before they converge on the correct hypothesis alone (see, for example, Hochberg (1988)).[4]

An alternative is that perhaps children place some belief in both hypotheses to begin with, and gradually alter their belief in which hypothesis is correct based on the ability of each hypothesis to account for the input data (Clark 1992, Legate and Yang 2007, Pearl 2009, Pearl 2011, Pearl and Lidz 2009, Pearl and Weinberg 2007, Yang 2002, Yang 2004). Parameter–setting here behaves more like a light dimmer, with the "switch" resting at a position between two options and its relative position indicating which option is more likely. If this is more similar to how children set their parameters, children's learning behavior can easily intermix different parametric options. The different options are simply accessed probabilistically, depending on the belief the child has in each option. For example, if a child places 60% probability in the language being extrametrical and 40% probability in the language not being extrametrical, this child might choose to produce linguistic data that use extrametricality 60% of the time and linguistic data that do not use extrametricality 40% of the time. As children are exposed to more input, they alter their beliefs in the different options until the correct option is most likely (a probability near 100%) and the incorrect option is not at all likely (a probability near 0%). Notably, this takes some time, during which the child will probabilistically access both hypotheses and so use both parameter values. This account thus seems to accord better with observable child language acquisition behavior.

3.2 Batch learning or incremental learning?

When children are updating their hypothesis about the correct parameter value, do they amass a quantity of data points and then update based on the inferences that come from those data as a group (batch learning) or do they update as the data come in (incremental learning)? The question of batch vs. incremental learning is really dealing with a larger issue in the language acquisition literature, specifically this: what question are we trying to answer about language acquisition?

The question of *learnability* is asking what is, in principle, possible to learn given the available data. This concerns what information may be useful given the available data and the knowledge that must be learned, but typically without considering constraints that humans have when learning the knowledge from the data. Models answering this question are called "ideal learner", "rational", or "computational–level" models (Foraker, Regier, Kheterpal, Perfors, and Tenenbaum 2009, Goldwater, Griffiths, and Johnson 2009, Hsu and Chater 2010, Perfors, Tenenbaum, and Regier 2006, 2011). Because these models assume no memory constraints and no processing constraints, they use batch learning (usually over an entire corpus's worth of data) and may use inference procedures that are unlikely to be used by humans.

The question of *acquirability* is asking what is possible to learn given the available data, and the constraints children have when they use that available data to learn. Models answering this question are sometimes called "algorithmic–level" models (Fodor and Sakas 2004, Gibson and Wexler 1994, Legate and Yang 2007, Niyogi and Berwick 1996, Pearl 2009, Pearl 2011, Pearl, Goldwater, and Steyvers 2010, Pearl, Goldwater, and Steyvers 2011, Pearl and Lidz 2009, Pearl and Mis 2011, Pearl and Weinberg 2007, Sakas 2003, Sakas and Fodor 2001, Sakas and Nishimoto 2002, Yang 2002, Yang 2004) referring to Marr's second level of representation for questions of information processing (Marr 1982). Because these models assume children have memory constraints, they often make the simplifying assumption that children process data as they are encountered (incremental learning), rather than storing detailed data for analysis later on. These models also typically assume children have processing constraints and try to use inference algorithms or inference algorithm approximations that are more likely to be feasible given these processing constraints.

If we are interested in explaining children's observable behavior, computational–level models can help us determine if the acquisition problem is solvable in principle when framed a particular way (ex: Can children learn the correct setting of the extrametricality parameter, given these data?). If the acquisition problem is not solvable even in principle, this suggests that something is amiss in the formulation of the acquisition problem – perhaps the knowledge to be attained is not what we think it is, or the child has additional restrictions on potential hypotheses that we are not considering (see Pearl (2011) for more discussion of this point for metrical phonology parameters). Once we determine that an acquisition problem is solvable in principle, we can then use algorithmic–level models to determine if it is solvable by constrained learners such as children. This allows us to explore what is required to make knowledge that is learnable in principle acquirable in practice.

3.3 Parameters: Ensembles or individuals?

Several researchers have viewed the acquisition problem as a search problem within the hypothesis space of possible grammars (Clark 1992, Gibson and Wexler 1994, Niyogi and Berwick 1996, Pearl 2009, Pearl 2011, Yang 2002) – which of these grammars are able to account for the data in the input, and which aren't? A learning algorithm like this is error–driven, with grammars treated as atomic units that either perform well on the data or perform poorly. While children recognize that grammars are comprised of parameters with particular values, and these values can be changed, the performance of the grammar as a whole is what counts. In essence, the child rewards a successful grammar and punishes an unsuccessful one.

Computational–modeling researchers have instantiated this idea in various ways. Clark (1992) imagines the child scoring grammars based on their relative "fitness"; grammars able to account for more data are viewed as more fit and rewarded, while grammars able to account for less data are less fit and punished. The child rewards and punishes grammars as a whole, rather than keeping track of the fitness of the parameters within the grammars. Gibson and Wexler (1994) and Niyogi and Berwick (1996) view the child as rewarding or punishing individual parameters within grammars, based on the grammars' ability to account for the data. If the current grammar cannot account for the current data point, the child may choose to consider a new grammar that differs from the old one by a single parameter value as long as this new grammar accounts for the data point in question. In this way, parameters that were not the problem do not get punished.

Yang (2002) and Pearl (2009, 2011) extend this idea by allowing parameter values to be rewarded or punished probabilistically, based on the performance of grammars using these values. Rather than getting rid of a parameter value if a new grammar without it can account for the data, the child simply lessens her confidence in that parameter by lessening her confidence in *all* the parameters involved in the unsuccessful grammar. Some parameter values are certainly unfairly punished this way, but the idea is that those unfairly punished will be part of successful grammars later on and be rewarded, while the real culprits will only be part of unsuccessful grammars, and so continually be punished. The upshot is that the child does not necessarily know which parameter was the culprit for any given failure of a grammar to account for data, but can still track how confident she is that any given parameter value is a good one for the language by the performance of grammars as a whole on the data she encounters. This can skirt the issue of blame assignment (sometimes known as the credit problem (Dresher 1999)) that arises for ambiguous data (see examples (5) and (11) above). Specifically, the child does not need to pinpoint which parameter is at fault in order to get information out of an ambiguous data point. She simply observes if the current parameter values under consideration are able to collectively analyze the data point, and updates her confidence in those parameter values accordingly.

This last part about tracking individual parameters (even indirectly) is quite useful because it helps reduce the hypothesis space for acquisition. To illustrate this, suppose a child knows there are 25 binary parameters. If the child views the problem as a search among grammars made up of these 25 parameters, there are 2^{25} (33,554,432) hypotheses. If instead the child views the problem as setting 25 parameters in one of two ways, the hypothesis space is more like 25*2 = 50. That is, there are 25 choices with 2 options each. Of course, these 25 choices can produce 2^{25} different grammars, but that is not what the child is explicitly tracking: She only cares which way each parameter is set. So, if the child focuses instead on what the choice is for each parameter, rather than on how grammars as a whole perform, the hypothesis space is much smaller.

3.4 Getting the most out of your data?

Several researchers have considered how children might explicitly track the values for specific parameters. Based on computational modeling results like Gibson and Wexler (1994) and Niyogi and Berwick (1996), researchers such as Fodor (1998), Dresher (1999), and Lightfoot (1999) propose that children are specifically keying into parts of the observable data that are linked to a specific parameter value. In particular, children are waiting for data that are unambiguous with respect to a given parameter. This requires that children already know what parts of the observable data are important for a given parameter, or can derive it for each parameter value in some way.[5] Still, once they have this knowledge, acquisition is much simpler. The main point is that the process of acquisition is inherently different for a child interested only in identifying the correct parameter values for her language: She is scouring the data for the relevant pieces, rather than worrying about which grammar can account for the most observed data.

Of course, this doesn't mean children would ignore the available data. Rather, they scour the data for the unambiguous "cues" (Dresher 1999, Lightfoot 1999) or designated structures (Fodor 1998), and reward parameter values that are best able to account for the distribution of these specific data. The difference is subtle, but important. A child interested in matching the data with an entire grammar will, naturally, choose the parameter values that belong to the grammar that fits the data the best. A child interested in tracking cues within the data for different parameter values may end up choosing a grammar comprised of parameter values that individually match the data the best, but which, when combined into a grammar, are less compatible with the data.

Let's consider an example from metrical phonology that demonstrates this from Pearl (2008). The English metrical phonology system is thought to have the rightmost syllable of the word be extrametrical (+extrametrical) (Dresher 1999). If a child tracks which grammar as a whole is compatible with the most English child–directed speech data, the answer is a grammar that has no extrametricality (-extrametricality). However, if a child is searching for unambiguous data regarding extrametricality, the cues for +extrametricality are more frequent than the cues for -extrametricality.[6] In this way, a child looking at overall grammar "fitness" can end up with a different answer than a child looking for unambiguous cues to parameter values in the data – and in this case, we think it is the child using unambiguous data who ends up with the right answer for English.

3.5 Who goes first?

A factor that appears when the child is considering parameters individually this way is the "learning path" (Dresher 1999, Lightfoot 1989, Lightfoot 1999, Baker 2005, Pearl 2008), which is the order in which the child learns the values of the parameters. If the child is comparing grammars as a whole, this does not matter – parameters function solely as part of a grammar, and the grammar either succeeds or doesn't. However, when a child is learning parameters values individually, there are a variety of orders that parameter values might be learned in. Because parameters interact, learning the value of one parameter may influence the child's interpretation of the data to come. Let's consider an example from before, (5) repeated (more briefly) here as (12).

(12) Analyses for "giRAFFE"

(a) +extrametrical (leftmost syllable), feet–headed–left:
Analysis: gi (RAFFE)
(b) +extrametrical (leftmost syllable), feet–headed–right:
Analysis: gi (RAFFE)
(c) -extrametrical, feet–headed–right:
Analysis: (gi RAFFE)

Suppose the child did not know whether the metrical feet in the language were headed on the left or headed on the right. Given that lack of knowledge, this data point is ambiguous for extrametricality, since there are analyses for both +extrametricality and -extrametricality. A child learning from unambiguous data therefore learns nothing from this data point. However, suppose that same child encountered this data point after learning that the language had metrical feet headed on the left. Then, the only analysis remaining is (12a), with +extrametricality. This same data point is now perceived as unambiguous for the extrametricality parameter, where before it was not.

The learning path turns out to be crucial for learning the English metrical phonology systems using parameters. Pearl (2008) discovered that a child tracking unambiguous data probabilities will decide on the correct parameter values for English, but only if the parameters are learned in particular orders. This has a direct implication for acquisition – not only do English children need to know the parameters and values these parameters can take, but they also need to know what order to learn parameters in. Fortunately, the learning path knowledge sometimes turns out to be derivable from properties like the distribution of data in the input (see Pearl (2007) for discussion). And perhaps this is a small price to pay for being able to reduce the hypothesis space by learning about parameters individually.

One reason why unambiguous data are so useful is that they provide maximally informative data. Since unambiguous data, by definition, are compatible with only one value for a parameter, they strongly signal that this parameter value is the right one. This leads to another good reason to consolidate current ideas about parameters, so that more structural properties about the language are connected. If structural properties are linked through a parameter, unambiguous data for one structural property serve as unambiguous data for all the rest of the structural properties linked by that parameter. For example, suppose our current conception of language parameters results in the following situation: parameter P1 has some unambiguous data while parameter P2 does not. This is problematic for an unambiguous learner. However, suppose that parameters P1 and P2 are really related, and so actually are both instances of parameter P*. Since unambiguous data exist for P1, and P1 tells us about P*, we then have unambiguous data for P2 (which is also an instance of P*). In general, consolidating parameters seems to be a good idea, as unambiguous data are not all that common. Pearl (2008) shows that the most frequent unambiguous data for parameters in the metrical phonology domain investigated made up less than 5% of the child's available input. In fact, one of the criticisms of the unambiguous data approach was that unambiguous data may not really exist for all parameters. At that point, a child learning only from unambiguous data would be in big trouble.

One solution would be to back off on the claim that children only learn from unambiguous data, under the assumption that unambiguous data really are non–existent (or at least very rare) for some parameters. While children probably should not learn from the data if said data are completely ambiguous, there may be many data points that are only somewhat ambiguous, rather than completely ambiguous. For example, consider again our metrical phonology analyses for "giRAFFE", repeated here as (13):

(13) Analyses for "giRAFFE"
(a) +extrametrical (leftmost syllable), feet–headed–left:
Analysis: gi (RAFFE)
(b) +extrametrical (leftmost syllable), feet–headed–right:
Analysis: gi (RAFFE)
(c) -extrametrical, feet–headed–right:
Analysis: (gi RAFFE)

A child considering these three analyses does not unambiguously know which value is correct for extrametricality and feet headedness. However, she still might be able to place some confidence in one value over the other based on the number of analyses that include one value vs. the other. In this case, since two of the three analyses involve +extrametricality, she might believe that +extrametricality is more likely than -extrametricality; similarly, since two of three analyses involve metrical feet headed on the right, she might believe right–headed metrical feet are more likely than left–headed metrical feet for the language. So, while an unambiguous learner would ignore this data point as uninformative, a child leveraging whatever information is available might find this data point partially informative. This means that even if no unambiguous data exist for a given parameter, some informative data may still exist. A child willing to learn from any informative data is not in nearly so tough a spot as a child waiting for unambiguous data (cf. Fodor and Sakas (2004) for a striking example of this in syntactic acquisition). Recent computational modeling research has shown that this is a powerful idea for learning structural properties such as the fact that language has hierarchical structure (Perfors, Tenenbaum, and Regier 2006, 2011) and the structure that the English referential element *one* has (Foraker et al. 2009, Pearl and Lidz 2009, Pearl and Mis 2011, Regier and Gahl 2004).

3.6 Insufficient data: Does not compute?

Still, there may be some cases where a language provides no unambiguous data and also has very little informative data of any kind. This can happen when a parameter is not used in a particular language. While all parameters may be available in all human minds, some may only come into play when certain other structural properties of the language hold. For example, Baker (2001) describes a parameter known as "adjective neutralization" which determines whether adjectives are treated like verbs or like nouns in the language. Importantly, this is only relevant for polysynthetic languages like Mohawk and Walpiri. Non–polysynthetic languages, like English, do not use this parameter. So, perhaps unsurprisingly, English does not include informative data that show adjective neutralization. Another example of a parameter with very little informative data in a given language is verb–raising in Korean.

Verb–raising concerns the position of the tensed verb with respect to adverbs/negative elements (Baker 2001, Cook and Newson 1996, Yang 2004, among others). When a clause has both tense (e.g., +present) and an adverb or negative element (e.g., *often*, *not*), languages vary on the position of the verb. In some languages, such as French, the verb "raises" or is "attracted" to the structural position of the tense, so that it ends up as a tensed verb preceding adverbs or negative elements (14a). In other languages, such as English, this does not happen, and the verb appears to follow adverbs or negative elements (14b). However, in a head–final language like Korean, very few data points distinguish whether Korean is +verb–raising or -verb–raising (14c), since the verb occurs in the same linear position independent of its structural height (see Han, Lidz, and Musolino (2007)).

(14) Verb–raising variation

(a) +verb–raising (French)

Underlying: Jean [+present] souvent/pas voir Marie
Jean *often /not* *see–infinitive* *Marie*

Observable: Jean voit souvent/pas Marie
Jean *see–present* *often* */not* *Marie*
“Jean often sees Marie”/ “Jean does not see Marie.”

(b) -verb–raising (English)

Underlying: John [+present] often see Mary.
Observable: John often sees Mary.
Underlying: John [+present] not see Mary.
Observable: John does not see Mary.

(c) ?verb–raising (Korean)

(i) Underlying: Yuri cacwu Toli–lil ttyali [–n–ta]
Yuri *often* *Toli*–acc *hit* [pres–decl]

(ii) Observable Yuri cacwu Toli–lul ttayli–n–ta
Yuri *often* *Toli*–acc *hit*–pres–decl
“Yuri often hits Toli.”

(iii) Analysis 1:
+verb–raising Yuri cacwu Toli–lul t_{ttayli} ttayli–n–ta
Yuri *often* *Toli*–acc t_{hit} *hit*–pres–decl
“Yuri often hits Toli.”

(iiv) Analysis 2:
-verb–raising Yuri cacwu Toli–lul ttayli–n–ta $t_{n–ta}$
Yuri *often* *Toli*–acc *hit*–pres–decl $t_{pres–decl}$
“Yuri often hits Toli.”

Except in very rare constructions, the effects of verb–raising are masked in Korean due to other structural properties of the language. Because of this, children and most adults may not encounter enough data to decide whether their language has verb–raising or not. More specifically, almost all data are compatible with both raising the verb and not raising the verb. What does this mean for Korean speakers? Do they simply pick one value in the absence of any informative data? Or do they leave it as undecided, so that if pressed to show knowledge of the parameter, they would flip flop between the two options with some probability? In the case of Korean, Han, Lidz, and Musolino (2007) found that there seemed to be two groups of adult speakers: those that raise their verb and those that do not. Very few speakers were willing to allow both options with some probability. So, this seems to indicate that Korean speakers had selected a parameter value despite having very little informative data. Still, perhaps the few data points they encountered in their lifetimes were enough to influence their choice – some statistical learning techniques are able to capitalize on very small amounts of data in the input (see Foraker et al. (2009) for an example of this). However, when Han, Lidz, and Musolino examined children's knowledge, children too seemed to split into a group that raises their verbs and a group that does not. It is possible that these children also may have all encountered just enough informative data, but the likelihood is much less. Instead, it may be that children really were choosing one parameter value in the absence of informative data. If so, this tells us something very interesting about how the human mind functions during language acquisition. Specifically, this suggests that the human mind prefers to make decisions and set a parameter – even for seemingly irrelevant parameters – rather than allow optionality to exist at this structural level of representation. That is, even if a parameter does not appear to be used in a language, the human mind prefers to set it than leave it unset.

An increasingly common way to explore how exactly the human mind learns with parameters is to use computational modeling. Computational modeling gives us very precise control over the language acquisition process, including what hypotheses the child entertains, what data are considered relevant, and how the child changes belief in different hypotheses based on the data. For learning with parameters, we can control which parameters are considered, and which values of those parameters are considered. We can also control if children learn only from unambiguous data, or consider other kinds of data informative. In addition, we can control if data are immediately influential, or if a child waits until she has some confidence that the data really do indicate a particular parameter value is correct. A variety of data are available from experimental studies that show us what a child's input data really look like, and we can ground our models with this empirical data to make them as much like a child's language acquisition process as possible. When we have a model like this, we can embody ideas like "learning only from unambiguous data" and "waiting until the data are really conclusive". If some of these ideas are required for the model to behave like children do, this tells us that these ideas may do a good job of describing how the human mind acquires language. For instance, if we find that, when given Korean data, one model ends up setting the verb–raising parameter while another model prefers to leave it unset, this suggests that the first model is a more accurate representation of the language acquisition process.

This last point is relevant for computational models that use statistical analyses such as Bayesian inference when compared to models that use update algorithms such as linear–

reward penalty (e.g., see work by Pearl (Pearl 2009, Pearl 2011) and Yang (Legate and Yang 2007, Yang 2002, Yang 2004)). Suppose a learner begins with two hypotheses, and gives each a probability of 0.5. A Bayesian inference model that receives no input will not update the probabilities since both hypotheses are still equally likely. In contrast, a learner using a variational learning update process (see discussion of this update process in section 4 below) will end up converging to one hypothesis or the other. This is due to the random fluctuations that are part of this style of update, with small random preferences being magnified over time. Still, it should be noted that very little data is much different than no data for a Bayesian learner, while this difference is not so dramatic for a linear reward–penalty learner. Thus, to decide between these two kinds of updating processes, it may matter very much that there is very little data that children could notice, as opposed to there actually being no informative data in the input at all.

4. Specific examples of learning with parameters

Let's now look at a few examples of computational models that investigate language acquisition using parameters. We will look within the domains of syntactic acquisition and metrical phonology, and find that the same considerations seem to arise about what the relevant data are and how the learning algorithm should work, even if the hypothesis space (consisting of particular parameters) is agreed upon. In addition, all these models share something besides using parameters to define the hypothesis space: They all seek to combine insights from statistical learning into a cognitively–plausible model.

For syntactic acquisition, work by Yang (2002, 2004) has examined the problem of learning the correct parameter values from the data children encounter, such as whether or not the language has verb–second movement and movement of wh–words to the front of the clause in questions. All of Yang's models use the frequencies of data found in child–directed speech samples. As mentioned in the previous section, Yang's models, often called "variational learners", do not specifically look for highly informative data; instead, they consider all data relevant for learning. The particular algorithm used by Yang incorporates a model from mathematical psychology called the linear reward–penalty scheme (Bush and Mosteller 1951). The algorithm recognizes that each parameter has two competing values (e.g., ±wh–fronting). Initially, the model associates a probability of 0.5 with each, representing its belief that each parameter value is equally likely. This probability is then altered, based on the data encountered.

For each data point, the variational learner generates a grammar based on the current probabilities associated with all parameter values. For instance, when generating the value for the wh–movement parameter, the model uses the probabilities associated with +wh–movement and -wh–movement. Suppose they are 0.40 and 0.60 respectively; then, the model will use the +wh–movement value with 40% probability and the -wh–movement value with 60% probability. This value selection process continues until all parameter values have been selected and a complete grammar is formed. Using the probabilistically generated grammar, the model then generates a structure for the data point (e.g., generating the structure for a question and the word order associated with that structure). If the generated word order matches the observed word order for the data point, all parameter values that were selected to participate in the grammar are rewarded; if the generated order does not match, all participating parameter values are punished. Notably, this model does not attempt to assign credit of blame to a particular parameter value within the grammar. Instead, all participating values are rewarded or punished together. The model then moves on to the next data point.

Updating (whether rewarding or punishing) is specified by the linear reward–penalty scheme. The update equation involves a parameter γ that determines how liberal the model is. The larger γ is, the more probability the model shifts for a single data point.

(15) Updating using the linear reward–penalty scheme

p_v = previous probability of parameter value

p_o = previous probability of opposing parameter value

(a) generated word order matches observed word order (reward)

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

$p_{o_{new}} = (1 - \gamma)p_o$[7]

(b) generated word order does not match observed word order (punish)

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

$p_{o_{new}} = \gamma + (1 - \gamma)p_o$

As an example, suppose we consider the probabilities of +wh–movement and -wh–movement for the wh–movement parameter. Initially, they are both 0.5. For the first data point, suppose -wh–movement is chosen to be part of the grammar and that grammar fails to generate the observed word order (e.g., the observed data point is “Who is that?” and the generated word order is “That is who?”). The -wh–movement value (and all other participating values) are punished. Suppose γ is 0.01. The new value of -wh–movement would be (1-0.01)*0.5 = 0.495 and the new value of +wh–movement would be 0.01 + (1-0.01)*0.5 = 0.505.

If the model rewards or punishes parameter values every time a data point is encountered, this is known as the Naïve Parameter Learner (Yang 2002). The idea, discussed in the previous section, is that the incorrect parameter values will be punished more often than the correct parameter values since the correct parameter values are the ones that should be compatible with the language data. Yang also advocates a variant of this model called "batch learning" that is more conservative about changing its beliefs – instead of updating after every data point, the model waits to update a parameter value until a string of failures or successes has been observed for that value. It does this by keeping count of how many successful and unsuccessful parses a parameter value has been involved in. So perhaps a more transparent name for it is a "counting learner". The benefit of a counting learner is that it smoothes the acquisition trajectory when parameters that interact are involved. Instead of jumping to conclusions after encountering ambiguous data, the counting learner waits for awhile to see if it's really true that a particular parameter value works or doesn't work.

Let's look at an example of this. Suppose we have a counting learner that has a count size of 5. If the count for a parameter value reaches 5, the parameter value is rewarded. If the count reaches -5, the parameter value is punished. Every time the parameter value is part of a grammar that generates a structure that matches the observed data point, the counter is increased; conversely, every time the parameter value is part of a grammar that generates a structure that mismatches, the counter is decreased. After the reward or punishment, the parameter value's counter is reset to 0. Suppose now we again consider the probabilities of the values for the wh–movement parameter. Initially, both + and -wh–movement have a probability of 0.5, and their counters are both 0. For the first data point, suppose +wh–movement is chosen to be part of the grammar and that grammar fails to generate the observed word order. The counter for +wh–movement is now -1. For the next three data points, suppose -wh–movement is chosen for the grammar and those grammars succeed at generating the observed word order. The counter for -wh–movement is +3 and the counter for +wh–movement is -1. Suppose the next two data points use +wh–movement and those grammars succeed: -wh–movement's counter is still +3, but +wh–movement's counter is now +1. Suppose then that the next six data points use +wh–movement and those grammars fail: -wh–movement's counter is still +3, but +wh–movement's counter is now -5, which is the count limit. The +wh–movement value is then punished using the appropriate update equation for the model. If the model uses a γ of 0.01, the new probability of +wh–movment is 0.495 and the new probability of -wh–movement is 0.505. The counter for +wh–movement is then reset to 0.

The conservativity of this learning model can be seen from the previous example – instead of updating for each of the twelve individual data points (punishing +wh–movement once, rewarding -wh–movement three times, rewarding +wh–movement two times, and then punishing +wh–movement six times), the model only punishes +wh–movement once. Importantly, this is only after the +wh–movement value has been involved in a string of failures, and so is more likely to really be failing.

Using this kind of learner, Yang discovered that a strong gauge of acquisition success was the quantity of unambiguous data for each parameter value in the input, even though his learners weren't specifically looking for just unambiguous data. The intuition is that ambiguous data for a parameter can be parsed by either parameter value, and so neither value is guaranteed to be punished (although either *might* be if some other parameter interacts and causes a mismatch with the data point). However, for unambiguous data, one parameter value is guaranteed to be punished every single time – by its very nature, an unambiguous data point is compatible with only one parameter value, no matter what other parameters are involved. Taking the idea that unambiguous data for parameters value is a key notion, Yang (2004) investigated the age of acquisition for a number of parameter values and the amount of unambiguous data available in child–directed speech for these parameter values. He found a strong correlation, where the more frequent unambiguous data was, the earlier a parameter value seemed to be acquired by children (16). This suggests that unambiguous data certainly are quite important for learning with parameters, even though they may not be the only data children use.

(16) Correlations between age of acquisition and unambiguous data frequency[8]

(a) wh–fronting in English questions (+wh–movement)

Example unambiguous data: “<u>Who</u> did you see?’

Unambiguous data frequency: 25% of input

Age of acquisition: As early as children can be tested

(b) verb–raising in French (+verb–raising)

Example unambiguous data type: Jean <u>voit souvent</u> Marie.

Jean sees often Marie
"Jean often sees Marie."

Unambiguous data frequency: 7% of input

Age of acquisition: 1 year, 8 months

(c) obligatory subject in English (-pro–drop)

Example unambiguous data: "There's a penguin on the ice."

Unambiguous data frequency: 1.2% of input

Age of acquisition: 3 years

(d) verb–second in German (+verb–second)

Example unambiguous data: Pinguine liebe ich.
Penguins like I.
"I like penguins."

Unambiguous data frequency: 1.2% of input

Age of acquisition: 3 years

(e) no medial–wh in English (-medial–wh)

Example unambiguous data type: "Who do you think is on the ice?"

Unambiguous data frequency: 0.2% of the input

Age of acquisition: after 4 years

A complementary line of computational work by Sakas and Fodor (Fodor 1998b, Fodor and Sakas 2004, Sakas 2003, Sakas and Fodor 2001, Sakas and Nishimoto 2002) investigates other learning strategies a child might have for acquisition, with particular concern for how much time it would take for a child to converge on the right grammar. They quantified the general notion of time as how many data points a child would need to observe. To provide their models with a realistic acquisition problem, they required their models to set 13 binary parameters that interact, for a hypothesis space of 3072 possible grammars. The data the models learned from represented the sentence types that could occur in the language, such as "Subject Verb" (a data point fitting this description in English would be *I laughed*). All of the strategies they investigated were also based on the idea previously discussed that the child is attempting to generate a structure that matches the observable word order. A summary of some of the different strategies is shown in (17) below. The learning strategies are divided into those that only learn when there is a failure to match the observed data (error–driven models (17a–c)), and those that are sensitive to whether the current data point is ambiguous for which grammar generated it (structural triggers models (17d–f)).

(33) Acquisition strategies

(a) Error–Driven Blind Guess: Only change the current guess for the grammar if that grammar cannot generate the observed word order. At this point, make a new guess randomly for what the entire grammar is.

(b) Trigger Learning Algorithm: Only change the current guess for the grammar if that grammar cannot generate the observed word order. At this point, optionally change some parameter values.

(c) Error–Driven Variational Learner: Only change the current guess for the grammar if that grammar cannot generate the observed word order. At this point, probabilistically generate a new grammar based on probabilities associated each parameter value. If the new grammar can generate the observed word order, reward all participating parameter values; otherwise, punish all participating parameter values.

(d) Strong Structural Triggers Learner: The learner is aware of all possible structures capable of generating the observed word order. Only parameter values required by all these structures are selected as the correct one for the language, as the data point is unambiguous for these values.

(e) Waiting Structural Triggers Learner: The learner is aware if more than one structure is capable of generating the observed word order, and where exactly the ambiguities in structure are. It only adopts parameter values that correspond to parts of the structure that occur before any ambiguities, and are therefore unambiguous.

(f) Guessing Structural Triggers Learner: The learner is aware if more than one structure is capable of generating the observed word order, and where exactly the ambiguities in structure are. For the ambiguous parts, it chooses one parameter value, based on some structural heuristic.

Sakas and Fodor find that error–driven learners generally fare much less well than learners that care about the ambiguity of the data. Specifically, the error–driven learners (17a–c) take longer to converge on the correct grammar for the language. However, among the learners that are sensitive to the ambiguity in the data, it turns out that there is a tradeoff between learners who only learn from unambiguous data and learners that are willing to make a guess even when the data are ambiguous. While the unambiguous learners (17d–e) will sometimes converge very quickly if the data distribution is favorable (that is, contains unambiguous data), they will take a very long time to converge if the data distribution is unfavorable. The less conservative learner (17f) may take longer on average to converge, but this learner type doesn't have the extremes in variation because it does not depend on unambiguous data being available. What this tells us is that unambiguous data may be very useful, but if a child is attempting to generate the data she encounters rather than identify cues for parameter values, she may need to rely on somewhat ambiguous data as well.

Another line of computational work in the metrical phonology domain by Pearl (Pearl 2007, Pearl 2008, Pearl 2009, Pearl 2011) examines an additional complication for the acquisition problem. In the models previously discussed, there has been very little noise in the input data – that is, there are very few misleading data points for the correct grammar of the language (which might occur as speaker errors, or as part of different dialects). However, in metrical phonology, this is not always the case. It may very well be that there is a set of data that are exceptional, but still part of the language. 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 lexical marking. When Pearl examined English metrical phonology data, she discovered that at least 27% of the English child–directed speech data were irregular with respect to the metrical analysis for English derived from Dresher (1999), Halle and Vergnaud (1987), and Halle and Idsardi (1995), again because of lexical exceptions. This makes acquisition a bit tougher than before – not only are the data often ambiguous, but some of them may be inconsistent with the target grammar because of lexically listed exceptions.

Pearl (2009, 2011) examined how well learners that probabilistically learn from all data would do on this dataset, since this seemed another realistic example of the data children encounter. The learning models examined were the variational learner from Yang (2002) and an incremental Bayesian model, both non–counting and counting versions. The variational model uses the update equations described in (15), while the incremental Bayesian model uses the update equations shown in (18).[9] If a parameter value participates in a grammar that generates a stress contour that matches the observed stress contour, the number of successes for that parameter value is incremented by 1. If a parameter value participates in a grammar that does not, the number of successes is left alone. Either way, the total data seen is incremented by 1 if the parameter value was part of the grammar used to generate the stress contour. The probabilities for opposing parameter values are then calculated and all probabilities are normalized so they sum to 1. So, for each parameter value, the model tracks (a) the current probability, (b) the number of matching stress contours that parameter value has been involved in generating, and (c) the total number of stress contours that parameter value has been involved in generating.

(18) Update equations for incremental Bayesian learning model

p_v = previous probability of parameter value

p_o = previous probability of opposing parameter value

$$p_{vnew} = \frac{1.5 + successes}{3 + total\ data\ seen}$$

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

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

As an example, suppose we look at the extrametricality parameter. Initially, the probabilities for both + and -extrametricality are 0.5. For the first data point, suppose +extrametricality is chosen to be part of the grammar and that grammar fails to generate the observed stress contour. The +extrametricality value (and all other participating values) are punished. The non–normalized probability for the +extrametricality value is (1.5+0)/(3+1) = 0.375. The non–normalized probability for the -extrametricality value has not changed from 0.5 since it was not used for this data point. The normalized probability of +extrametricality is then 0.375/(0.375 + 0.5) = 0.429 while the normalized probability of -extrametricality is then 0.571.

Pearl found that none of these learners succeeded with any kind of reliability, perhaps a somewhat disheartening discovery. The trouble turned out to be that the English grammar as a whole was compatible with less data than other grammars available in the hypothesis space, even though the individual parameter values may have been compatible with more data. This is again the problem of tracking overall grammar "fitness" with the data compared to identifying the parameter values involved in the system, based on cues from the data. If we believe children do end up choosing the parameter values in the English grammar, they must have some way of viewing the acquisition problem such that these parameter values are the most optimal ones for explaining the data that are relevant.

One idea is that children are sensitive to the unambiguous data, and the ability of the parameter values to account for the unambiguous data are what matters – not the overall fitness with the entire dataset. Pearl (2008) found that children would choose the English grammar (and more particularly, the parameter values that make up the English grammar) if they specifically track the unambiguous data distributions. This was far more heartening news for acquisition for two reasons. First, it showed that unambiguous data do indeed exist in a very realistic acquisition scenario, which is always a concern if we believe children rely on them. Second, it showed that they appeared in the correct distributions to lead a child to the right parameter values. The reason why learning from unambiguous data worked is because the unambiguous data favor the English parameter values when the parameters are acquired in particular orders. So, if the parameters are acquired in one of those orders, the English parameter values are the fittest for the unambiguous data. In that case, a probabilistic learning algorithm that prefers the optimal values (as most do) will converge on the English grammar. Interestingly, additional simulations suggested that it is not the ordering alone that causes the English parameter values to be optimal. When parameters were set in similar orders by the models that learn from all data (both ambiguous and unambiguous), there was no reliable convergence on the English grammar. Among the learners that do converge, there seemed to be no common learning path. This suggested that the culprit was the ambiguous data. Learning from these data (though granted in a rather naïve way) will mislead an English learner. These simulations also have led to a testable prediction about children's acquisition trajectory: If children are using parameters and relying on unambiguous data, they must learn parameter values in certain orders. This prediction remains to be tested, but the key point is that it has a given us something to look for in children's behavior that we didn't know to look for before.

These three sets of modeling studies have all investigated the acquisition of language using parameters in different ways. Though their general strategies of investigation differed somewhat, they all demonstrated both how useful unambiguous data can be and also some of its potential pitfalls for a parameter–setting learner. We have also seen how computational modeling not only assesses different learning strategies but also how it can generate testable predictions about acquisition.

5. Conclusion

In this chapter, we have discussed some ideas for what linguistic parameters are intended to be, highlighting their similarity to statistical parameters and describing how and why they would be beneficial for language acquisition. We have tried to emphasize that using parameters to characterize the space of possible languages does not in itself provide a model of learning. In addition, a learning theory needs a mechanism for drawing inferences from observed data to particular parameter values. A number of alternative approaches to parameter–setting highlighted the learning choices that remain, such as why learning by using individual parameters might be better than learning parametrically–defined grammars that are treated as atomic units, and whether and how learners rely on strictly unambiguous data. We subsequently reviewed several computational modeling studies to demonstrate the contribution computational modeling can make to the parameter–setting enterprise. Specifically, we can use computational modeling techniques to explore what learning strategies make a parametric grammar acquirable. If we find certain parametric grammars are not acquirable from available child input, this can bear on the validity of the proposed parameters that comprise those grammars (cf. Frank and Kapur 1996). Even with the correct parameters, children may still need to bring additional learning biases to the task of language acquisition in order to correctly set their parameters – another possibility that can be explored easily within a computational modeling framework. One of the benefits of explicit computational modeling approaches to parameter–setting is that alternatives may differ in the predictions they make about the time course of acquisition (e.g., based on the amount of data required to set the parameters as in Yang (2004)), which makes it possible to test the validity of parametric approaches against the behavior of actual learners. We hope that this chapter inspires continued research into the linguistic parameters that could make acquisition as swift and relatively easy as it seems to be.

NOTES

[1] Note that this is ungrammatical in English, which does show a *that*–trace effect.

[2] Metrical feet are indicated by parentheses (…), while stress is indicated by CAPITAL LETTERS henceforth.

[3] This is true even if it takes some time for children to choose the correct parameter value (perhaps because they will not change their current hypothesis unless they are sure that it is really unable to analyze the input). No matter when they decide to choose the correct parameter value, once they choose it, they should continue to use it exclusively.

[4] Note, however, that the "light switch" account may be able to produce more gradual–seeming behavior if it used for learning entire grammars rather than individual parameter values. See section 3.3 for more discussion of learning grammars vs. learning individual parameters.

[5] Fodor (1998) and Lightfoot (1999) suggest that children might be able to discover the relevant unambiguous cues by a process they already use for comprehension: parsing the incoming data. To "parse" a data point is to know the underlying structure for a data point, i.e., to know what parameter values generated it. If a particular parameter value is required for a successful parse of a given data point, that data point is unambiguous for that parameter value.

[6] This is true provided that the value for the extrametricality parameter is learned after the

value of the quantity sensitivity parameter. See discussion below regarding the learning path.

[7] This can also be calculated as 1- p_{vnew} since there are only two parameter values available. The same is true for the punishment calculation.

[8] The underlined portion of the example unambiguous data represents the most relevant portion of the data point.

[9] We note that the equation for p_{vnew} includes numbers in the numerator (1.5) and denominator (3) that represent the learner bringing minimal prior knowledge about which parameter value is correct, believing in both equally at the start. This learner is strongly driven by the data points encountered. See Pearl (2011) for a more detailed discussion of the intuitions behind this equation.