

To appear in *Experimental Methods in Language Acquisition Research*.

## Using computational modeling in language acquisition research

Lisa Pearl\*

### 1. Introduction

Language acquisition research is often concerned with questions of *what*, *when*, and *how* – what children know, when they know it, and how they learn it.

Theoretical research traditionally yields the *what* – the knowledge that children attain. For instance, this includes how many vowel phonemes the language has, how the plural is formed, and if the verb comes before or after the object. These and many other questions must be answered before the child can speak the language natively. This linguistic knowledge is the child's goal.

Experimental work traditionally provides the *when* – at what point in development the child attains particular knowledge about the language. Of course, there is a certain logical trajectory. It would be difficult to discover how the past tense is formed before being able to identify individual words in fluent speech. Still, this logical trajectory does not offer precise ages of acquisition. Experimental work can, for example, pinpoint when word segmentation occurs reliably and when English children correctly produce past tense forms. This gives us the time course of language acquisition. The child can segment words reliably by *this* age, and apply regular past tense morphology by *that* age, and so on.

Then, there is the *how* – how the child learns the appropriate *what* by the appropriate *when*. This is the mechanism of language acquisition, which includes what knowledge is required to reach the adult knowledge state at the appropriate time. Computational modeling can be used to examine a variety of questions about the language acquisition process, because a model is meant to be a simulation of the relevant parts of a child's acquisition mechanism. In a model, we can precisely manipulate some part of the mechanism and see the results on acquisition. If we believe the model accurately reflects the child's language acquisition mechanism, these manipulations and their effects inform us about the nature of that mechanism. Importantly, some manipulations we can do within a model are difficult to do with children. The modeling data are thus particularly useful because of the difficulty of getting those same data through experimental means.

Within a model, we can choose the hypotheses the child entertains. For example, consider the syntactic structure that generates the observable word order of a language, such as *Subject Verb Object*. Should a child only entertain hypotheses that are hierarchical (i.e. they involve clustering words into larger units like phrases)? Or, could the child also consider linear hypotheses (where words of a sentence are viewed as a single large group that has no special divisions within it)? This definition of the child's hypothesis space would be hard to implement in a traditional experiment because, while we can assess what hypotheses a child is entertaining (e.g. see Crain & Nakayama (1987)), we cannot easily *control* the hypotheses a child has about the pattern of data presented.

Within a model, we may also constrain the data children learn from. Though the *input* consists of all available data in the linguistic environment, the child's data *intake* may or may not include all these data. For example, consider word order. There are many languages that seem to alter the basic word order of the language in certain linguistic contexts. For German, many theoreticians believe the basic order is *Subject Object Verb*. However, the word order in main clauses is often *Subject Verb Object*, which could be generated by movement options in the grammar. If a child is trying to decide the basic order of the language, Verb-Object or Object-Verb, should the child only use data that unambiguously signal one option? Or, should the child

use all available data, and guess between the two when the data are ambiguous? As with the hypothesis space definition, this kind of data intake definition is also hard to implement in a traditional experiment. If we believe children only need a subset of the available data to acquire the adult language successfully, the logical experiment would be to give children only the restricted input set and then see if they acquire the adult language correctly. If they need more than that subset, their acquisition will be derailed. However, we cannot simply lock children up in a room for a few years, only allow them to hear various subsets of data from their native language, and then see the effect on their acquisition. It is unethical (and a logistical nightmare besides). However, this is precisely what we can do with our modeled child. If the modeled child with the restriction is successful while the modeled child without the restriction is *not*, we have reason to believe that children may filter their input to the relevant data subset.

Within a model, we can also alter how children use data to update their beliefs in various hypotheses. Turning again to word order, suppose the child has encountered a datum signaling Verb-Object order. Should this immediately increase the likelihood of the Verb-Object hypothesis? Or, should the child wait until she encounters more Verb-Object data, in case this datum was some kind of fluke? If the child *does* update her beliefs based on this datum, how much should they be updated? This kind of manipulation, like the others discussed above, is simply not feasible to implement experimentally, as we cannot easily control how children change their beliefs. Modeling, however, provides a way to manipulate this.

Modeling's strength is its ability to create a language acquisition mechanism we have complete control over. In this way, we garner data we could not easily get otherwise. However, the point of modeling is to increase our knowledge about the way *human* language acquisition works, not simply provide a model capable of solving a particular problem. We must thus be careful to ground our model empirically – i.e., we must consider if the details of the model are psychologically plausible by looking at the data available on human language acquisition from theoretical and experimental research. We should remember that modeling is an additional tool we use to understand language acquisition, not a replacement for others we already have.

Of course, despite good intentions, most models in the real world may not satisfy all psychological plausibility criteria. This is the difference between modeling ideals and modeling reality. In practice, the real test of a model is whether it reveals something interesting we did not know before and whether it generates testable predictions. Of course, it is easier for a model to do both of these when the model is empirically grounded.

## **2. Rationale**

We generally model to answer questions about the nature of language acquisition that we cannot easily test otherwise. But exactly what questions are these?

It is quite useful to step back momentarily, and think about how to characterize the general problem of language acquisition. Marr (1982:24-29) identified three levels at which an information-processing problem can be characterized: (a) the computational, which describes what the problem is, (b) the algorithmic, which describes the steps needed to carry out the solution, and (c) the implementational, which describes how the algorithm is instantiated in the available medium. Marr's insight was that these three levels are distinct and can be explored separately. Even if we do not understand how the solution can be implemented, we can know what the problem is and what considerations a psychologically plausible algorithm needs to have. Moreover, understanding the problem at one level can inform the understanding of the problem at other levels.

This transfers readily to language acquisition. We can identify the computational-level problems to be solved: stress assignment, word segmentation, word order rules, etc. A psychologically plausible algorithm should include considerations like the available memory

resources children have, and how much processing is needed to identify useful data. The medium where all solutions must be implemented is the brain. Crucially, we do not need to know exactly how a given algorithm is instantiated in neural tissue. Consider stress assignment as a specific example. We can identify that the algorithm must involve processing and assigning stress to syllables, without knowing how neurons translate sound waves into the mental representation of syllables.

Note that the levels are not completely disconnected from each other. Knowledge of the algorithmic level, for instance, can constrain the implementational level for stress assignment. If we know the solution involves recognizing syllables within words, we can look for neural implementations that can recognize syllables.

For language acquisition, we can ask questions at all three levels. At the computational level, we can identify the problem to be solved, including definitions of both the input and the output. For our stress assignment example, the input is the available data in the linguistic environment, organized into syllables. The output is syllables with a certain amount of stress assigned to them. At the algorithmic level, we can identify psychologically plausible algorithms that allow the child to learn the necessary information from the available data. With stress assignment, considerations may include what linguistic units probabilistic learning should operate over (syllables, bisyllable clusters, metrical feet, etc.). At the implementational level, we can test the capability of biologically faithful models for implementing given algorithms and producing solutions that are behaviorally faithful. Neural networks are an example of biologically-inspired models that attempt to replicate human behavior in this way, as is the framework ACT-R (Anderson 1993).

Models are used to provide insight for problems that are not readily solvable. Testing the obvious with a model will, unsurprisingly, give obvious answers. For example, suppose, we have a model that learns the word order of verbs and objects in the language. A question inappropriate for modeling might be to ask if the model will always learn Verb-Object order when given examples of only Verb-Object order. Unless the model incorporates some very strong biases for another word order, the model will of course learn Verb-Object order. The model's output is unsurprising. No serious question will have been answered by this model.

Similarly, modeling does not provide informative answers to uninformative questions. A good rubric of informativeness is theoretical grounding. An example of an uninformative question is to ask if the model will hypothesize that the past tense is formed by not changing the word form when its input consists only of words ending in *-yze* (e.g. *analyze*) and *-ect* (e.g. *protect*). This is uninformative because there is no theoretical grounding - no particular behavior from the model will yield anything more about the problem. Whether the model does or does not hypothesize the no-change past tense behavior, it is unclear what information we have gained. Without a theory that makes predictions one way or the other, all we have done by modeling this question is practice our computer programming skills.

In short, a model provides a way to investigate a specific claim about language acquisition, which will involve a non-obvious informative question. An example informative question might involve testing an acquisition theory that claims children should not learn from all the available data in order to acquire the correct generalizations about the language. Instead, children should only learn from "good" data, where "good" is defined by the acquisition theory. If a model is provided with data from the language and incorporates the theory's "good" data bias, will the model learn the correct generalizations about the language at the same rate children do?

Obviously, this is a very abstract question that can be instantiated numerous ways. One instantiation can be found in a study of learning word order by Pearl & Weinberg (2007), where children learned whether their language was *Verb Object* or *Object Verb*. There, a learning theory by Lightfoot (1991) claimed that children should learn only from word order data in main clauses (as opposed to data in embedded clauses). Moreover, children should learn only from

data perceived as unambiguous for a particular word order (Lightfoot 1999). Unambiguous data are compatible only with one hypothesis, while ambiguous data are compatible with more than one hypothesis. For example, unambiguous data for *Verb Object* would be compatible only with the Verb-Object order and not the Object-Verb order. To implement their model, Pearl & Weinberg used this acquisition theory to define the abstract notion of “good” data as unambiguous word order data found in main clauses.

The question mentioned above is informative for several reasons. First, the question is grounded theoretically in a claim about the data children use during acquisition. Second, the model can be grounded empirically from language data and the time course of acquisition that come from experimental work. Third, the model provides an informative test of the theory’s prediction. If the model learns the correct generalizations at the same rate children do, then the theory’s “good” data bias is supported. However, if the model does not display the correct behavior, then the theory’s claim is considerably weakened as it does not succeed when tested explicitly. For these reasons, this model’s behavior is both non-obvious and informative – and so the question is good to model.

Regarding the details of model implementation, empirical grounding is key. This can include using realistic data as input, measuring the model’s learning behavior against children’s learning behavior, and incorporating psychologically plausible algorithms into the model. These all combine to ensure that the model is actually about acquisition, rather than simply about what behavior a computational algorithm is capable of producing.

Let us examine word segmentation in detail as an example. Realistic data would be child-directed speech, which would be the un-segmented utterances a child is likely to hear early in life. These data can come from transcripts of caretakers interacting with very young children. An excellent resource for this kind of data is the freely available Child Language Data Exchange System (CHILDES) (MacWhinney 2000).

Measuring the model’s learning behavior against known acquisition behavior would include being able to segment words as well as children do and being able to learn the correct segmentations at the same rate that children do. Both of these measures – the correct segmentations and the correct rate of learning to segment - will come from experimental work that probes children’s word segmentation performance over time.

Psychologically plausible algorithms will include features like gradual learning, robustness to noise in the data, and learning incrementally. A gradual learner will slowly alter its behavior based on data, rather than making sudden leaps in performance. A robust learner will not be thrown off when there is noise in the data, such as slips of the tongue or chance data from a non-native speaker. An incremental learner is one that learns from data points as they are encountered, rather than remembering all data encountered and analyzing it altogether later. These features are derived from what is known about the learning abilities of children – specifically, what their word segmentation performance looks like over time (it is gradual, and not thrown off by noisy data) and what cognitive constraints they may have at specific ages (such as memory or attention limitations).

Without this empirical grounding – without realistic data, without measuring behavior against children’s behavior, and without psychological plausibility considerations – the model is not as informative about how humans learn. Since language acquisition is about how humans learn, models should be empirically grounded as much as possible if they are to have explanatory power.

Yet, we should not go too far in empirically grounding the model – no model can include everything about a child’s mind and linguistic experience. It is simply not feasible to do so. The crucial decisions in modeling involve where to simplify. A model, for instance, may assume that children will pay equal attention to each data point encountered. In real life, this is not likely to be true – there are many factors in a child’s life that may intervene. Perhaps the child is tired or distracted by some interesting object nearby. In these cases, the data at that point in time will

likely not impact the child's hypotheses as much as other data have or will. Yet it would be an unusual model that included a random noise factor of this kind.

The reason for this excision is that unless there is an extremely pervasive pattern in the noise due to varying levels of attention in the child, the model's overall behavior is unlikely to be affected by this variable. Generally, a model should include only as many variables as it needs to explain the resultant behavior pattern. If too many parameters of the model vary simultaneously, the cause of the model's behavior is unknown – and so there is less explanatory power.

The solution, of course, is very similar to that of more traditional experimental work: isolate the relevant variables as much as possible. The key word is relevant. It is alright to have some model parameters that vary freely or only have their value fixed after their effect on the model's behavior is assessed. For example, the input to the model is a certain number of data points, and that quantity may need to be set only after observing its effect on the model's behavior. The modeler should always assess the effect the value for a model parameter has on the model's behavior. For the input set size, does the behavior change if the model receives more data points? If so, then this is a relevant parameter after all. Does the behavior remain stable so long as the input quantity is above a certain number? If so, then this is only a relevant parameter if the input size is below that threshold. In explaining the model's behavior, this input size variable can be removed as long as its value exceeds that critical threshold.

A good general strategy with free parameters in a model is to systematically vary them and see if the model's behavior changes. If it doesn't, then they are truly irrelevant parameters – they are simply required because a model needs to be fully fleshed out (for instance, how much input the model will encounter). However, these parameters are not part of the real cause of the model's behavior. Still, if the behavior is dependent on the free parameters having some specific values or range of values, then these become relevant. In fact, they may become predictions of the model. For instance, if the model only performs appropriately when the input quantity is greater than the amount of data encountered by a child in 6 months, then the model predicts that this behavior should emerge later than 6 months after the onset of acquisition.

It is reasonable to ask why models have free parameters, instead of only including parameters specified by the theoretical claim the model is investigating. The reason is that theoretical claims are rarely as fleshed out as a model needs to be. They may not say exactly how much data the child should encounter; they may not predict the exact time of acquisition or even the general time course; they will often make no claims about how exactly children update their hypotheses based on the available data. These (and many others) are decisions left to the modeler.

Variables common to most models include how much data the model processes and the parameters involved in updating the model's beliefs (usually in the form of some equation that requires one or more parameters, such as the equations involved in the algorithms mentioned in the next paragraph). The input to the model can usually be estimated from the time course of acquisition. Suppose a child solves a particular learning task within 6 months; the amount of data a child would hear in 6 months can be estimated from transcripts of child-directed speech.

The update of the model's beliefs usually involve probabilistic learning of some kind, which in turn involves using some particular algorithm. Three examples of algorithm types are those used in Linear reward-penalty (Bush & Mosteller 1951, used in Yang 2002, among others), neural networks (Rumelhart & McClelland 1986, Plunkett & Marchman 1991, Hare & Elman 1995, Plunkett & Juola 1999, among others), and Bayesian updating (used in Perfors, Tenenbaum & Regier 2006, Pearl & Weinberg 2007, Pearl & Lidz *submitted*, among others). No matter the method, it will involve some parameters (Linear reward-penalty: learning rate; neural networks: architecture of network; Bayesian updating: priors on hypothesis space). It is alright to have free parameters in the model, but it is the modeler's responsibility to (a) assess their effect on the model's behavior, and in some cases (b) highlight that these are instrumental to the model's behavior and are therefore predictions the model makes about human behavior.

Three ways to evaluate a model's contribution to language acquisition are its *formal sufficiency*, *developmental compatibility*, and *explanatory power*. Formal sufficiency asks if the model learns what it is supposed to when it is supposed to from the data it is supposed to. This is evaluated against known child behavior and input. Developmental compatibility asks if the model learns in a psychologically plausible way, using resources and algorithms the way a child could. This is evaluated against what is known about a child's cognitive capabilities. Explanatory power asks what the crucial part of the model is for generating the correct behavior, and how that impacts the theoretical claim the model is testing. This is evaluated by the modeler via manipulation of the model's relevant parameters. When these questions can be answered satisfactorily, the model contributes something significant to language acquisition research.

### 3. Linguistic variables

Simply speaking, modeling can be applied to any acquisition problem where there is a theoretical claim, a defined input set, and a defined output behavior. This can range from identifying phonemes to word segmentation to learning word order rules to identifying the correct parameter values for complex linguistic systems. This section surveys a number of modeling studies for a variety of language acquisition tasks. In each case, the model's strength is in its empirical grounding and its ability to make testable predictions. Because we obviously cannot include all relevant studies, the interested reader is encouraged to look within the studies mentioned for references to additional modeling studies examining similar acquisition problems.

#### 3.1. Aspects of the sound system

Modeling can be applied to the problem of discovering the phonemes of a language. Vallabha, McClelland, Pons, Werker, and Amano (2007) investigated the acquisition of vowel contrasts in both English and Japanese from English and Japanese vowel sound data. The acquisition task was well-defined: can a model learn the relevant vowel contrasts for these languages without explicit knowledge about the relevant dimensions of variation and the number of distinct vowels? This task is non-trivial, especially since the model receives no explicit feedback regarding the correctness of its hypotheses. The data came from English and Japanese mothers speaking to their children, and so were a realistic estimation of the data children encounter. The learning algorithms were incremental variants of probabilistic algorithms from computer science. The model was fairly successful, depending on the type of learning algorithm used. One implication for acquisition was that learning probabilistically from noisy data can lead to human-like performance, even without defining the hypothesis space very strictly. Moreover, the type of probabilistic learning significantly influences how successful acquisition is. A prediction from this model might be that the processes underlying acquisition are more similar to the more successful algorithm – in this case, perhaps involving an assumption about how the acoustic data are generated.

Modeling can also be used to investigate the acquisition of metrical phonology, a complex linguistic system that determines where the stress is in words (Dresher & Kaye 1990, Dresher 1999, Pearl 2008). For instance, the word *emphasis* has stress only on the first syllable 'em': it is pronounced EMphasis. Generative metrical theory believes that this stress pattern is generated by a system that groups syllables into larger units called metrical feet, and a number of parameters describe how the grouping works. Languages vary on how they group syllables, and so vary on what values these parameters have. The child's task is to unconsciously infer the parameter values that lead to the stress patterns observed in the input.

Pearl (2008) examined this acquisition problem for English, which has many exceptions to the general rules of the language. Child-directed English speech from the CHILDES database

was used as input, and the measure of successful acquisition was whether the English parameter values could be learned from these data. This model specifically tested a claim that children can only succeed if they learn exclusively from unambiguous data (Dresher 1999, Lightfoot 1999). As an example of unambiguous data in this model, consider that one parameter was whether all syllables are included in metrical feet. Unambiguous data for the English value are compatible only with an analysis that does not include all syllables in metrical feet; ambiguous data are compatible both with an analysis where all syllables are not included *and* with one where all syllables are included.

The results showed that children with a bias to learn only from unambiguous data could succeed. In addition, acquisition success was only guaranteed if the parameter values were learned in a particular order. A prediction generated from this model is that English children should learn the English parameter values in that special order if they really are learning only from unambiguous data.

### 3.2. Aspects of words

Another problem modeling is used for is understanding how children extract the units we think of as words from fluent speech, i.e. word segmentation. Experimental work on artificial languages suggests that infants can unconsciously track the statistical information known as *transitional probability* between syllables, e.g. the probability for syllable sequence AB that syllable B is next when syllable A is the first syllable. One question is if this strategy succeeds on realistic data.

Gambell and Yang (2006) modeled the performance of a transitional probability learner on English child-directed speech. The data came from transcripts of English caretakers speaking to children, drawing from the speech samples available in CHILDES. To transform the written transcripts into the sounds children hear, Gambell and Yang used a freely available pronunciation dictionary, the CMU Pronouncing Dictionary (<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>), that transforms written words into individual sounds. For example, the word “eight” would be transformed into the sound sequence “EY T”, which contains two sounds (as opposed to five letters).

It turns out that a transitional probability learner actually performs quite poorly on the English dataset. Further exploration by Gambell and Yang showed that when a transitional probability learner is armed with additional information about the sound pattern of words (specifically, an assumption of one primary stress per word), the modeled learner succeeds. Interestingly, this assumption yields success even if the learner does not use transitional probabilities. A prediction from this model is that this knowledge is very useful to have, and we can then test if children have it before they can segment words. Because this model was explicitly defined, the learning procedure could be precisely manipulated and informative predictions made about strategies children might use to solve this task.

Another task modeling can investigate is the grammatical categorization of words. Grammatical category information tells the child how the word is used in the language – for instance, nouns (but not verbs) can be modified by adjectives: *juicy peach* (but not *juicy eat*). Wang & Mintz (2008), building on work by Mintz (2003), explored one strategy children might use to identify words that behave similarly: *frequent frames*.

Frequent frames consist of framing words that cooccur frequently in the child’s input. For example, in *she eats it*, the frame is *she \_\_\_ it* for the word *eats*. This strategy was motivated by experimental evidence suggesting that infants can track the cooccurrence of items that are non-adjacent. Frequent frames were intended as a means to initially cluster similarly behaving words in languages with relatively fixed word order. Notably, frames do not rely on word meaning, unlike some other theories of grammatical categorization.

The data used as input for the model came from transcripts of child-directed speech from

CHILDES. The modeling demonstrated that a frequent frame learner can indeed successfully identify words that behave similarly solely on the basis of their common frames. The resulting categories mapped well to the “true” grammatical categories like noun and verb. However, note that not all words belonging to a particular grammatical category were identified as being in the category, e.g. not all nouns were grouped into the noun category (see section 6.1 for discussion). This implies that, while useful for languages with fixed word order, frequent frames cannot be solely responsible for children’s grammatical categorization. A prediction generated from this model was that children are sensitive to the information in frequent frames when learning a word’s grammatical category. Experimental work by Mintz (2006) tested the proposed sensitivity in 12-month-olds, and found that they do seem to use this distributional information.

Modeling can also be applied to learning morphology. One problem commonly examined is acquisition of the English past tense, due to the English data resources available and the potential impact on larger questions in language acquisition. The problem itself is one of mapping: given a verb (*blink, sing, think*), map that form to the appropriate past tense form (*blinked, sang, thought*). The input to models is usually realistic estimates of the verbs children encounter during acquisition, derived from resources like CHILDES. The output of the model is compared against what is known from experimental work about how and when children learn certain past tense forms.

The main point of interest in many morphology models is that there is a division between a regular pattern and several irregular patterns (e.g. *blink-blinked* vs. *sing-sang, think-thought* in the English past tense). Experimental work indicates that many English children have a trajectory that involves good performance on all the verbs they know, followed by poor performance on only the irregular verbs, which is then followed by good performance on all the verbs again. The ability to generate this learning trajectory (good-poor-good performance) can be one output goal for English past tense models. Another goal can be to assess if the correct behavior can result without the model explicitly learning a regular rule (e.g. *+ed* in the English past tense).

The learning procedures of these models usually try to consider psychological plausibility with some seriousness, and often vary between neural networks (Rumelhart & McClelland 1986, Plunkett & Marchman 1991, Prasada & Pinker 1993, Hare & Elman 1995, Plunkett & Juola 1999, Nakisa, Plunkett, & Hahn 2000, among others) and probabilistic rule-learning models (Yang 2002, Albright & Hayes 2003, Yang 2005, among others). Most models are incremental, learning as the data come in. When the models are able to produce the correct output behavior, it is because of some precise design feature within the model – perhaps the order data are presented to the model (e.g., Rumelhart & McClelland 1986) or what causes the child to posit a regular rule pattern (e.g., Yang 2005).

Of course, all these models make assumptions about the knowledge available to children. For instance, they assume that children know the underlying form of a word when they encounter the surface form (e.g. the child knows *thought* is the past tense of *think*), which may not be true in real life. As mentioned in the rationale section, these are simplifying assumptions on the part of the modeler. However, even simplified models can offer good insights into language acquisition with respect to what will (and will not) work, given the best possible acquisition scenario.

The predictions generated from these models pertain to the causal factors of the output behavior. For instance, the model by Yang (2005) predicts that the performance trajectory depends very precisely on the number of regular and irregular verbs encountered by the child and the order in which these verbs are encountered. This prediction can be assessed by examining specific input and performance data from experimental work with children learning the English past tense, and seeing if the model’s predictions match children’s behavior.

### 3.3. Aspects of syntax and semantics

Modeling can also be used to investigate the acquisition of syntactic and semantic

representations, and the connection between them. This is necessary for referential linguistic elements, such as anaphors, pronouns, and other referring expressions. An interesting property of referential items is they are only interpretable if the listener knows what they refer to. For example, the word *one* in English can be used referentially (known as *anaphoric one*): “Jack has a red ball - he wants another one.” Most adult English speakers interpret this to mean “He wants another *red ball*.” Thus, the word *one* refers to the words *red ball* (not just *ball*), and the referent of *one* in the world is a ball that is red (not just any ball). The correct interpretation of *one* relies on identifying the words *one* refers to (*red ball*), which then leads to the object in the world *one* refers to (a ball that is red). The problem for English children is acquiring this correct interpretation.

Several models have attempted to tackle this problem, using incremental, probabilistic learning algorithms on the data. Regier and Gahl (2004) and Pearl and Lidz (*submitted*) manipulated the data children use as input in their models, and found that the correct interpretation can be learned very quickly if children use only a highly informative subset of the available input. Foraker, Regier, Khetarpal, Perfors, & Tenenbaum (to appear) created a model that learned what words *one* referred to (e.g. *red ball* vs. *ball*) separately and prior to learning what object in the world *one* referred to (e.g. a ball that is red vs. any ball). While the models differ on their specifics, the general prediction is that children should be sensitive to specific aspects of the available data when acquiring this interpretation rule – and importantly, not learn from *all* available data. As before, because the hypothesis space and input to these models were precisely defined, the models could manipulate both and see the results on acquisition.

Modeling is also useful for examining the acquisition of word order rules in syntax. One example involves the formation of yes/no questions in English when the subject is complex. For instance, consider this sentence: “*The knight who can defeat the dragon will save the princess.*” The yes/no question equivalent is “*Will the knight who can defeat the dragon save the princess?*” Importantly, the auxiliary verb (*will, can, etc.*) that moves to the beginning of the question is the auxiliary verb from the main clause of the sentence (*The knight...will save the princess.*).

Interestingly, though children know this rule fairly early, the data they encounter have very few explicit examples of this rule – few enough that children’s early acquisition of it may seem surprising if their hypotheses for possible rules are not constrained (Legate & Yang 2002). However, given children’s statistical learning capabilities, Reali and Christiansen (2005) questioned whether a probabilistically learning child could infer the correct rule from simpler yes/no questions that are more abundant in the input. They designed a model sensitive to certain simple statistical information, called *bigrams*, that children might plausibly track in the data. A bigram probability refers to how often two words cooccur together in sequence. In the sentence “*She ate the peach*”, the bigrams are *she ate*, *ate the*, and *the peach*. Based on the input data (derived from CHILDES), a bigram model preferred the correct complex yes/no question over an incorrect alternative.

However, Kam, Stoynezhka, Tornyova, Sakas, and Fodor (2005), worried that this model’s success was due to particular statistical coincidences in the specific dataset used as input, and would not generally perform well. When they tried the bigram model on other datasets of child-directed speech, they found the model was at chance performance when choosing between yes/no question options. A prediction from these two models is that children must be learning the yes/no question formation rule from something besides bigram probability.

Other models have continued to examine this question (e.g. Perfors, Tenenbaum, & Regier 2006), as it relates to the knowledge children require to acquire language successfully. Put simply, if the information about the correct rule is available statistically in the data and children can access that statistical information, they do not require other prior knowledge to lead them to the correct rule.

Another type of syntactic modeling work concerns parametric systems popular in generative linguistic theory (e.g. Gibson & Wexler (1994), Niyogi & Berwick (1996), Sakas &

Fodor (2001), and Yang (2002)). One difficulty of parametric systems is interacting parameters, which makes identifying the parameter values responsible for an observable word order non-trivial. For instance, suppose a child hears a sentence with the form *Subject Verb Object*. Suppose also that the child was aware of two parameters: Verb-Object/Object-Verb (OV/VO) order and Verb-Second (V2) Movement (whether the Verb moves to the second position of the clause and some other phrase moves to the first position). The sentence mentioned could be due to different combinations of these parameters: (1) VO, no V2 (*Subject Verb Object*), (2) VO, V2 (*Subject Verb t<sub>Subject</sub> t<sub>Verb</sub> Object*), or (3) OV, V2 (*Subject Verb t<sub>Subject</sub> Object t<sub>Verb</sub>*). The goal of these models is to converge on the correct parameter values of the language, given the data available in the language. Yang (2002), in particular, considers the relative frequency of the different data types available to a child.

Each model's results demonstrate what is necessary to ensure children end up with the right parameter values. For example, the model in Yang (2002) demonstrates that children can learn from all data, so long as they use a probabilistic update procedure when converging on the correct parameter values. More generally, this model also provided a way to bridge the gap between acquisition via linguistic parameters and the empirical data that showed children's syntactic development was gradual. Traditionally, acquisition via linguistic parameters was believed to be necessarily abrupt - rather than gradual - which was problematic when trying to reconcile with the available empirical data. This model, however, produced a gradual trajectory by means of its probabilistic update procedure.

#### 4. Subjects

In modeling, the question is what kind of subject the model is *of*. All the modeling studies mentioned in section 3 used simulated learners who were normal monolingual (L1) speakers learning from monolingual data. However, modeling can be extended to other scenarios when the appropriate input data are available.

For example, we could create a second-language (L2) learning model that learns from L2 data. However, in contrast to an L1 model, the L2 model will already have linguistic information in place from its own L1. Importantly, we should ground the model theoretically and empirically. Theoretical grounding includes a description of the knowledge L2 learners have of their L1, how it is represented, and how this representation is altered or augmented by data from the L2 language. Empirical grounding includes the data learners have as input and what information they use to interpret that input (e.g., bias from their L1).

Similarly, the age of the simulated learner can vary. It is usually set at the age when the knowledge in question is thought to be acquired – information available from experimental work. For instance, in the Gambell and Yang (2006) word segmentation model, the simulated learner was assumed to be around 8 months. The age restriction in a model can be instantiated as the model having access to the data children of that age have access to (in the word segmentation case, syllables), and processing the data in ways children of that age would be able to process it (in the word segmentation case, without access to word meaning).

More generally, modeling different kinds of subjects requires a detailed instantiation of the relevant aspects of those subjects (e.g. knowledge known and initial bias). If this information can be reasonably estimated, an acquisition model can be designed for that subject. The key to an informative model is considering what the relevant information about the subject is and representing it in the model.

#### 5. Description of procedure

For modeling, the relevant experimental procedure is the model itself. Often, models are more concrete than the theories they test. This is both a strength and a weakness. A model's concreteness is good because it allows us to identify the aspects a theory may be vague about, e.g. how much data children process before learning the relevant information and how quickly children alter their linguistic knowledge when learning. The not-so-good part is that the modeler is forced to estimate reasonable values for these unknown parameters.

Most crucial is the decision process behind a model's design, not the details of how to program it. For this reason, we focus on the kinds of decisions that are most relevant for language acquisition models. All these decisions involve how the model will represent both the learner and the acquisition process. As theories often do not specify all the details a modeler needs to implement the model, the modeler must rely on other information sources to make the necessary decisions, e.g. experimental data and electronic databases like CHILDES. Still, the modeler's ingenuity is required to successfully integrate the available information into the model's design.

### *5.1. The effect of parameter values*

Sometimes, parameter values for a model can be estimated from available experimental data. For instance, the amount of data a child processes might be roughly equivalent to the amount of data the child has encountered by whatever age that knowledge is acquired. Other times, the modeler must choose a value for convenience and see if this strongly impacts the model's results. The learning rate in the model, for example, usually requires a value for specifying how much a single datum impacts the child's current hypotheses.

Since these parameters affect the outcome of the acquisition model, the value of these parameters may matter. We can check by trying a range of values for the unknown parameters and seeing the effect on the model. If the model's behavior is invariant, then these parameters, while necessary for implementing the model, do not really affect acquisition. In contrast, if the model only succeeds when the parameters have certain values, then this is a prediction the model makes about the actual values of these parameters in the acquisition process. For example, if the model only matches children's behavior when it receives more than a certain quantity of input, then the model predicts children need to encounter at least that much data before successfully acquiring the knowledge in question.

### *5.2. Control conditions and experimental conditions*

From a certain perspective, models are similar to traditional experimental techniques that require a control condition and an experimental condition so that the results can be compared. In modeling, this can correspond to trying ranges of parameter values for parameters that are not specified by the theory being tested. If the same results are obtained no matter what the conditions, then the variables tested – that is, the parameter values chosen for the model – do not affect the model's results.

Also, models that simulate children's ability to generalize can more transparently have control and test conditions. Suppose a model simulates children's ability to categorize sounds into phonemes, as in Vallabha *et al.* (2007). The model first learns from data in the input, e.g. individual sounds from child-directed speech. To gauge the model's ability to generalize correctly, the model must then be tested. The sound category model may be given a sound as input and then asked to output the category that sound belongs to. The control condition would give the model sounds that were in its input – i.e. sounds the model has encountered and learned from. The model's ability to correctly classify these sounds is its baseline performance. The test condition would then give the model sounds that were *not* in its input – i.e. these are sounds that the model has not previously encountered. Its ability to correctly classify them will demonstrate

whether it has correctly generalized its linguistic knowledge (as children do), or if it is simply good at classifying familiar data.

As we recall, data for models often comes from child-directed speech databases. Test condition data may come from a different speaker within that database. If the model has not learned to generalize, the model may perform well on data from one set of speakers (perhaps similar to the data it learned from) but fail on data from other speakers. This was the case for the word order rule model proposed by Reali & Christiansen (2005). While it was successful when tested on one dataset, Kam *et al.* (2005) showed that it failed when tested on another dataset. This suggests that the model is probably not a good reflection of how children learn since they can learn from many different data types and still learn the correct generalizations.

This last point is particularly important for models that import learning procedures (usually statistical) from more applied domains in computer science. Many statistical procedures are very good at maximizing the predictability of the data used to learn, but fail to generalize beyond those data. It is wise for a model using one of these procedures to show good performance on a variety of datasets, which underscores the model's ability to generalize. Since this is a property children's acquisition has, a model able to generalize will be more informative about the main questions in acquisition.

### *5.3. More practical details*

In general, a model will require a computer capable of running whatever program the model is built in. Sometimes, the program will be a software package where the modeler can simply input values for relevant variables and run it on the computer. For example, the PRAAT framework (Boersma 1999) functions this way, allowing a modeler to test the learnability of sound systems using a particular algorithm.

In general, however, modelers need to write the program that implements the necessary algorithm and describes the relevant details of the simulated learner. For this, a working knowledge of a programming language is vital – some useful ones that offer great flexibility are Perl, Java/C++, and Lisp. Often, it will not take a large amount of programming to implement the desired model in a particular programming language. The trickier part is the design of the model itself.

Modelers must consider what should be represented in the simulated learner, such as (a) how the model represents the required information (e.g. syllables or individual sounds), (b) if there is access to additional information during acquisition (e.g. stress contours of words during word segmentation), (c) how the model interprets data (e.g. if the model should separate words into syllables), and (d) how the models learns (e.g. tracking transitional probabilities between syllables). Again, theories are not usually explicit about all these details – but a model must be. Therefore, modelers will often spend a while making decisions about these questions before ever writing a single line of programming code.

## **6. Analysis and outcomes**

There are numerous ways to present modeling results, depending on what the model is testing. Unsurprisingly, the most effective measure for a model depends on the nature of the model, i.e. on what acquisition task it is simulating. The key is to identify the purpose of the model, and then present the results in such a way that they can be easily compared to the relevant behavior in children. Below, we review some common methods of representing modeling results.

### *6.1. Models that extract information*

For models that extract information, the relevant results are (not surprisingly) how well that information is extracted. Two useful measures, taken from computational linguistics, are *recall* and *precision*. To illustrate these two measurements, consider the task of a search engine like Google. Google's job is to identify web pages of interest when given a search term (e.g. "1980s fantasy movies"). The ideal search engine returns *all* and *only* the relevant web pages for a given term. If the search engine returns all the relevant web pages, its recall will be perfect. If the search engine returns only relevant web pages, its precision will be perfect. Usually, there is a tradeoff between these two measurements. A search engine can achieve perfect recall by returning all the web pages on the internet - however, only a small fraction of these web pages will be relevant, so the precision is low. Conversely, the search engine might return only a single relevant web page - precision is perfect (all returned pages were relevant), but recall is low because presumably there are many more relevant web pages than simply that one. Both precision and recall are therefore relevant for tasks of this nature, and both should be reported.

To transfer this to some models already discussed, consider Gambell and Yang's (2006) word segmentation model. Given a stream of syllables, the model tries to extract all and only the relevant words using different learning algorithms. Precision is calculated by dividing the number of real words posited by the number of total words posited. Recall is calculated by dividing the number of real words posited by the total number of real words that *should* have been posited. Often, the more successful strategies have fairly balanced precision and recall scores.

Another example is the word categorization model of Wang & Mintz (2008). Given a stream of words, the model clusters words appearing in similar frequent frames. These clusters are compared against real grammatical categories (e.g. verb) to see how well they match, with a given cluster assigned to a given grammatical category (e.g., cluster 23 is verb). Precision is calculated by dividing the number of words falling in that grammatical category within the cluster (e.g. all the verbs in the cluster) by the total number of words in the cluster. Recall is calculated by dividing the number of words falling in that grammatical category within the cluster (e.g. all the verbs in the cluster) by the total number of that grammatical category in the dataset (e.g. all the verbs in the corpus). Often precision is nearly perfect – but recall is very low. This implies frequent frames are very accurate in their classifications, but not very complete in classifying all the words that should be classified a particular way.

### 6.2. Models that match children's performance

Some models simulate the trajectory of children's performance – their results are the model's performance over time. This can then be matched against children's performance over time. For example, models of English past tense acquisition will often try to generate the "U-shaped" performance curve observed in children (e.g. Rumelhart & McClelland 1986, Yang 2005, among others). Specifically, the model aims to show an initial period where performance on producing verb past tenses is high (many correct forms), followed by a period where performance is low (usually due to overregularized forms like *goed*), followed again by a period where the performance is high. A successful model generates this trajectory without having the trajectory explicitly programmed in. The model explains children's behavior by whatever factor within the model generated this acquisition trajectory.

### 6.3. Models that reach a certain knowledge state

Some models measure how often acquisition succeeds within the model. For instance, the goal of Vallabha *et al.* (2007) was to correctly cluster individual sounds into larger language-specific perceptual categories. Different algorithms were tested multiple times and measured by how often they correctly classified a high proportion of individual sounds. The algorithm with a

higher success rate was deemed more desirable. This measurement generally demonstrates the robustness of the acquisition method. Ideally, we want a method that succeeds all the time, since (nearly) all children succeed at acquisition.

#### 6.4. Models that generalize

Some models measure how often a correct generalization is made. The models of Reali & Christiansen (2005), Kam *et al.* (2005), and Perfors *et al.* (2006) learned how to form yes/no questions (e.g. *Can the girl who is in the Labyrinth find her brother?*) from child-directed speech. The test was if the model preferred the correct way of forming a yes/no question over an incorrect alternative. If the model had generalized correctly from its training data, it would prefer the correct yes/no question all the time. As with the previous measurement, this measurement demonstrates the robustness of the learning method. If the model chooses the correct option all the time, it can be said to have acquired the correct generalization.

### 7. Advantages and disadvantages

Although every model is different, we can still discuss the main advantages and disadvantages of modeling without getting into the details of individual models.

The main advantage is the ability to precisely manipulate the language acquisition process and see the results of that manipulation. Generally, the manipulation should be something difficult to do with traditional experimental techniques. For example, it would be difficult for a traditional experiment to control the hypotheses children entertain, the interpretive biases they impose on the data, or the update procedure they use to shift belief between competing hypotheses. Modeling provides an effective way to test proposals related to these aspects of the acquisition mechanism.

The main disadvantage is that we can never be absolutely sure our model is really showing how acquisition works in children's minds. Perhaps some crucial information has been left out of the model's knowledge. Perhaps some critical oversimplifications have been made about how the model interprets the available data. Perhaps the output of the model lacks the nuances that children's behavior has. This is why empirical grounding is key. The more checkpoints on the model, the more we can believe what the model shows us about acquisition. This is where drawing from the results of experimental work can help.

In general, there is a dovetailing between experimental work and modeling studies. Experimental work can sometimes provide the empirical scaffolding a model needs to get off the ground. In return, models can sometimes provide predictions of behavior that can then be tested experimentally (e.g. Pearl (2008)). In this way, experimental research and modeling research continue to inform each other.

### 8. Dos and don'ts

Do:

- Read history: Learn from previous models about reasonable estimates of input, algorithms, and measures of output. Consider the strengths and weaknesses of prior models when designing your own.
- Listen to linguists: Linguists can provide the theoretical basis for the hypothesis space, and offer empirical data to base the model upon.
- Listen to psychologists: Psychologists will also provide empirical data to ground the model.

- Listen to computational linguists: Computational linguists will provide learning algorithms that can be implemented and adapted to be psychologically plausible as necessary.

Don't:

- Model when it is obvious: Models of obvious questions are not informative.
- Forget the theoretical and empirical grounding: Models that fail to use available data (both theoretical and experimental) as checkpoints are not as persuasive.
- Overlook that this is a model of *human* language acquisition: Psychological plausibility should be considered.

## Notes

\* This chapter was inspired in large part by Charles Yang's 2007 EMLAR lecture, and I am very grateful for his encouragement and insightful descriptions. In addition, I would like to thank the editors and two anonymous reviewers for very sensible suggestions. All views expressed in this chapter – insightful, sensible, or otherwise - are my own, however.

## References

- Albright, A. & Hayes, B. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90: 119-161.
- Anderson, J. 1993. *Rules of the Mind*. Hillsdale, NJ: Erlbaum.
- Boersma, P. 1999. Optimality-Theoretic learning in the PRAAT program. *Institute of Phonetic Sciences Proceedings*, 23: 17-35.
- Bush, R. R., & Mosteller, F. 1951. A mathematical model for simple learning. *Psychological Review*, 58: 313-323.
- Crain, S. & Nakayama, M. 1987. Structure dependence in grammar formation. *Language*, 63: 522-543.
- Dresher, E. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30: 27-67.
- Dresher, E. & Kaye, J. 1990. A computational learning model for metrical phonology. *Cognition*, 34: 137-195.
- Foraker, S., Regier, T., Khetarpal, A., Perfors, A., and Tenenbaum, J. To appear. Indirect evidence and the poverty of the stimulus: The case of anaphoric "one". *Cognitive Science*.
- Gambell, T. & Yang, C. 2006. Word segmentation: Quick but not dirty. Ms. Yale University.
- Gibson, E. & Wexler, K. 1994. Triggers. *Linguistic Inquiry*, 25: 407-454.
- Hare, M. & Elman, J. 1995. Learning and morphological change. *Cognition*, 56: 61-98.

- Kam, X., Stoyneshka, I., Tornyova, L., Fodor, J. D. & Sakas, W. 2005. Statistics vs. UG in language acquisition: Does a bigram analysis predict auxiliary inversion? In *Proceedings of the Second Workshop on Psycho-computational Models of Human Language Acquisition, Association of Computational Linguistics*, 69-71. Ann Arbor, MI: Association for Computational Linguistics.
- Legate, J. & Yang, C. 2002. Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*, 19: 151-162.
- Lightfoot, D. 1991. *How to Set Parameters: Arguments from Language Change*, Cambridge, MA: The MIT Press.
- 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.
- Marr, D. 1982. *Vision*. San Francisco: W.H. Freeman.
- Mintz, T. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90: 91-117.
- Mintz, T. 2006. Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. Golinkoff (eds), *Action Meets Word: How Children Learn Verbs*, 31-63. New York: Oxford University Press.
- Nakisa, R. C., Plunkett, K. & Hahn, U. 2000. Single and dual-route models of inflectional morphology. In P. Broeder & J. Murre (eds) *Models of Language Acquisition: Inductive and Deductive Approaches*, 201-222. Oxford: Oxford University Press.
- Niyogi, P. & Berwick, R. 1996. A language learning model for finite parameter spaces. *Cognition*, 61: 161-193.
- Pearl, L. 2008. Putting the emphasis on unambiguous: The feasibility of data filtering for learning English metrical phonology. In *BUCLD 32: Proceedings of the 32nd Annual Boston Conference on Child Language Development*, 390-401. Boston: Cascadilla Press.
- Pearl, L. & Lidz, J. Submitted. When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. Ms. University of California, Irvine & University of Maryland.
- 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 *28th Annual Conference of the Cognitive Science Society*. Vancouver, British Columbia: Cognitive Science Society.
- Plunkett, K. & Juola, P. 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science*, 23(4): 463-490.

- Plunkett, K. & Marchman, V. 1991. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38: 43-102.
- Prasada, S. & Pinker, S. 1993. Similarity-based and rule-based generalizations in inflectional morphology, *Language and Cognitive Processes*, 8: 1-56.
- Reali, F. & Christiansen, M. 2005. Uncovering the richness of the stimulus: Structural dependence and indirect statistical evidence. *Cognitive Science*, 29: 1007-1028.
- Regier, T. & Gahl, S. 2004. Learning the unlearnable: The role of missing evidence. *Cognition*, 93: 147-155.
- Rumelhart, D. & McClelland, J. 1986. On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP Research Group (eds), *Parallel distributed processing: Explorations in the microstructures of cognition*. Vol.2, *Psychological and biological models*, 216-271. Cambridge, MA: MIT Press.
- Sakas, W. & Fodor, J. 2001. The structural triggers learner. In S. Bertolo (ed.) *Language Acquisition and Learnability*, 172-233. Cambridge, UK: Cambridge University Press,
- 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.
- Wang, H. & Mintz, T. 2008. A dynamic learning model for categorizing words using frames. In *BUCLD 32: Proceedings of the 32nd Annual Boston Conference on Child Language Development*, 525-536. Boston: Cascadilla Press.
- Yang, C. 2002. *Knowledge and Learning in Natural Language*. New York: Oxford University Press.
- Yang, C. 2005. On productivity. *Yearbook of Language Variation*, 5: 333-370.