Computational Models of Language Acquisition
Lisa Pearl*

Abstract

This chapter describes the purpose of computational modeling in a language acquisition research domain. It discusses the types of questions computational modeling is ideally equipped to handle, reviews recent informative modeling studies for a variety of linguistic questions, and highlights important considerations for designing language acquisition models.

0. Introduction

Language acquisition research is often concerned with questions of *what, when*, and *how*. What do children know? When do they know it? How do they learn it?

Theoretical research describes the *what* – the knowledge that adult speakers have. How many vowel phonemes are there in the language? How is the plural formed? Does the verb come 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 or output state.

Experimental work often provides the *when* – at what point in time the child knows certain pieces of knowledge about the language and particular systems as a whole. The child follows a certain logical trajectory, of course. It would be difficult to discover how the past tense of verbs is formed before being able to identify individual words in fluent speech. Still, this logical trajectory does not offer any precise ages of acquisition. Experimental work can, for example, pinpoint when word segmentation occurs reliably and when English children learn the regular rule for the past tense. 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, we come to the *how*. How does the child learn the appropriate *what* by the appropriate *when*? This is the mechanism of language acquisition. Modeling work can be used for examining a variety of questions involving the language acquisition process. This is because a model is meant to be a simulation of the relevant parts of a child's language learning mechanism. In a model, we can manipulate some part of the mechanism very precisely and see the results on learning. If we believe the model has captured the salient aspects of a child's language acquisition mechanism, then these manipulations and their effects on learning inform us about the nature of that mechanism. Importantly, some manipulations that we can do within a model may be difficult to do with real children. So, the data we glean through modeling is uniquely informative because it would often be exceedingly tricky to get the same data through traditional experimental means.

0.1 What We Can Control

Within a model, we may choose the hypotheses the child considers for a particular problem. For example, consider the structure of language that generates the observable word order. Should a child only entertain hypotheses that are hierarchical (that is, 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 very hard to

* 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. All views expressed in this chapter – insightful or otherwise - are my own, however.

implement in a traditional experimental setup – how could we control the ideas the child has about the pattern of data presented?

Within a model, we may also constrain the data children use for learning. Though the *input* consists of all available data in the linguistic environment, the child's data *intake* may or may not include all of that 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. In German, many theoreticians believe the basic order is Subject Object Verb. However, the word order in main clauses is often Subject Verb Object, which is believed to 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 the available data, and simply 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. We can't 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's unethical (and a logistical nightmare besides). But this is, in effect, precisely what we can do with our modeled child.

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 data signaling Verb-Object order. Should this immediately increase the likelihood of the Verb-Object order hypothesis? Or, should the child wait until she has seen more Verb-Object data (what if this data was some kind of fluke)? If the child *does* update her beliefs based on this data point, how much should they be updated? This kind of manipulation, like the others discussed above, is simply not feasible to implement experimentally. How can we control the way children change their beliefs? Once again, modeling provides a way to manipulate this variable in the language acquisition mechanism.

0.2 What We Should Consider

Modeling's strength is its ability to create a language acquisition mechanism that we have complete control over. In this way, we garner data that we could not get otherwise. However, the point of modeling is to increase our knowledge about the way that *human* language acquisition works, not simply provide a computational or mathematical model capable of solving a particular problem. So, we must be careful to ground our model empirically – that is, we must consider if the details of the model are psychologically plausible by looking at the data available on human language acquisition. Are the hypothesis space, the data set, and the update procedures realistic? To inform us about how to implement our model, we rely on theoretical work about the nature of language and experimental work about children's knowledge of language. We can't design a realistic model without these. Modeling is an additional tool we use to understand language acquisition - not a replacement for others we already have.

1. Rationale

Now, why do we model? We model to answer questions about the nature of language acquisition that we can't easily test otherwise. But what questions *are* these exactly?

It's 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:

(1) Marr levels of description
  (a) computational level: describes what the problem to be solved is

(b) algorithmic level: describes the steps needed to carry out the solution

(c) implementational level: describes how the algorithm is instantiated in the available medium

The insight of Marr was that these three levels are distinct and can be explored separately. Even if we don't 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: phoneme identification, word segmentation, word order rules, stress contour rules, etc. A psychologically plausible algorithm will need to 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 don't need to know how exactly a psychologically plausible algorithm is instantiated in neural tissue. Let's take 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.

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

Speaking more generally about the language acquisition problem, we can ask questions at all three levels. At the computational level, we can identify the problem to be solved – which includes 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 psychologically plausible algorithms and producing solutions that are behaviorally faithful. Neural networks are a prime example of biologically inspired models that attempt to replicate human behavior in this way.

1.1. Some Modeling Caveats: Types of Questions

A model is meant to provide insight to problems that are not readily solvable. Testing the obvious with a model will, unsurprisingly, give obvious answers. Suppose for example that we have a model that learns the word order of verbs and objects in the language. A question *inappropriate* for modeling would be something like the following: "If I give the model examples only of Verb-Object order, will the model always learn Verb-Object order?" Unless the model incorporates some very strong biases for another word order, the model will of course learn Verb-Object order. So, the output of the model in this case is unsurprising. No serious question will have been answered by a model of this kind.

Similarly, modeling doesn't provide informative answers to uninformative questions. A good rubric of informativity is theoretical grounding. An uninformative question might be something like this: "If the model's input consists only of words ending with –yze (like *analyze*) and words ending with –ect (like *protect*), will it hypothesize that the past tense is formed by not changing the word form (*analyze* becomes *analyze, protect* becomes *protect*)?" This is uninformative because there is no theoretical grounding - no particular behavior from the model

will tell you anything more about the problem. If the model doesn't hypothesize the no-change past tense behavior, what will this tell you? If the model *does* hypothesize that behavior, what will *that* tell you? 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.

The main point is that a model provides a way to investigate a specific claim about language acquisition, which will involve a non-obvious informative question. Here's an example of one: Suppose that a language learning theory claims that a child shouldn't learn from all the available data in order to learn the correct generalizations about the language. Instead, the child should only learn from "good data", where "good" is defined by the learning theory. If a model is provided with data from the language and incorporates the learning theory's "good data" bias, will the model learn the correct generalizations about the language at the same rate children do?

This question is grounded theoretically in a claim about the data children use during acquisition. The model is grounded empirically from language data that comes from experimental work (such as child-directed speech transcripts) and from the time course of language learning that also comes from experimental work. Moreover, the model provides an informative test of the learning theory's prediction. If the model learns the correct generalizations at the same rate children do, then the learning theory's "really good data" bias is supported. On the flip side, if the model does not display the correct behavior, then the learning theory's claim is considerably weakened as it does not succeed in a realistic learning scenario. For these reasons, this model's behavior is both non-obvious and informative. Therefore, this kind of question is good to model.

1.2. Some Modeling Caveats: Empirical Grounding

Regarding the details of model implementation, empirical grounding is vital. 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. All of these combine to ensure that the model is actually about human learning, rather than simply about what behavior a computational algorithm is capable of producing.

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

Measuring the model's learning behavior against child learning 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 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 word segmentation learner is one that learns from data as it is encountered, rather than remembering all data encountered and analyzing it all 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 again children's behavior, and without a psychologically plausible model – the model is not informative about how humans learn. Since the point of language acquisition research is how

humans learn, computational 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's simply not tractable 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 the case – there are many factors in a child's life that may intervene. Perhaps the child is tired or is distracted from the speaker by some interesting object in the environment. In these cases, the data at that point in time will likely not impact the child's hypotheses about linguistic knowledge as much as other data has 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 to the "attention" noise of the child, the model's overall behavior is unlikely to be affected by this variable. In general, a model should include only as many variables as it needs to explain the resultant behavior pattern. Too many points of variation will cause the model to again lose explanatory power. Put simply, if too many parameters of the model vary simultaneously, the model's behavior cannot be attributed precisely to the manipulation of these variables. The cause of the model's behavior is unknown – and so there is no 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's alright to have some model parameters that vary freely or need to be calibrated. For instance, the input set to the model is a certain number of data points, and may not be specified explicitly by the learning theory. The important thing is to assess the effect the value for these additional model parameters 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 size 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 it's above 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). But these parameters are not part of the real cause of the model's behavior. However, 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 about the real state of the world. If, for instance, the model only performs appropriately if the input size 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 learning.

1.3. Some Modeling Caveats: Free Parameters Within the Model

Why do models have these free parameters, anyway? Why not just include only the parameters specified by the theoretical claim the model is investigating? This would be fine if theoretical claims about language acquisition were fully fleshed out to the extent that a model needs to be. Unfortunately, they rarely are. 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, and they will often make no claims about how exactly the child updates his linguistic knowledge based on the available data. These (and many others) are all decisions left to the modeler.

Variables common to most models are how much data the model processes and whatever parameters are involved in updating the model's beliefs (usually in the form of some equation

that requires one or more parameters). 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 popular algorithms are Linear reward-penalty (Bush & Mosteller 1951, used in Yang 2002, among others), neural networks (Rumelhart & McClelland 1986, among others), and Bayesian updating (used in Perfors, Tenenbaum & Regier 2006, Pearl & Weinberg 2007, among many 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). Again, it's alright to have free parameters in the model – it's simply up to the modeler 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.

1.4. Some Basic Questions for Modeling: Summary

There are three main types of questions for evaluating a model's contribution to language acquisition: questions of *formal sufficiency*, *developmental compatibility*, and *explanatory power*. First, formal sufficiency: does the model learn what it's supposed to learn when it's supposed to learn it from the data it's supposed to learn it from? This is evaluated against known child behavior and input. Second, developmental compatibility: does the model learn in a psychologically plausible way, using resources and deploying learning algorithms in a way a child plausibly could? This is evaluated against what's known about a child's cognitive capabilities. Third, explanatory power: what's the crucial part of the model that makes it work, and what does this mean for 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.

2. Linguistic Variables

Here we'll examine more closely the kinds of problems modeling can be applied to. To put it simply, modeling can be used for any linguistic problem where there is a theoretical claim about learnability, a defined input set, and a defined output behavior. This can range from identifying phonemes in the language to extracting words from fluent speech to learning word order rules to identifying the correct parametric system values for complex linguistic systems. In the remainder of this section, we'll survey a number of modeling studies for a wide variety of language acquisition tasks.

2.1. Phoneme Acquisition

First, modeling can be applied to the problem of discovering the phonemes of a language. Vallabha, McClelland, Pons, Werker, and Amano (2007) did just this, investigating the learnability of vowel contrasts in both English and Japanese from English and Japanese vowel sound data. The learning 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? 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 tried within the model were incremental and probabilistic, drawing from similar

algorithms in computer science. The performance of this model was quite promising – high success rates, depending on the type of learning algorithm used. The implications for learning theory are straightforward: learning probabilistically from noisy data can lead to human-like performance in this case, even without defining the hypothesis space very strictly. However, the type of probabilistic learning has a significant effect on how successful learning will be. A prediction from this model might be that the statistical processes underlying human learning will be similar to those properties of the algorithm that most closely matches human behavior.

2.2. Word Segmentation

Modeling can be applied to the problem of how children extract the units we think of as words from fluent speech. Experimental work on artificial languages suggests that certain kinds of statistical information can be tracked unconsciously by young children – namely, the *transitional probability* between syllables. The transitional probability of the syllable sequence AB is the probability that B is the next syllable, given that A is the first syllable. One question is if this strategy will be effective 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 CHILDES (MacWhinney 2000). Of course, the child encounters the pronunciations of these utterances, not their written form. So, Gambell and Yang used a freely available pronunciation dictionary, the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), to approximate the sounds children heard.

It turns out, perhaps surprisingly, 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 (called the Unique Stress Constraint), the modeled learner succeeds. Moreover, the Unique Stress Constraint yields success even if the learner doesn't use transitional probabilities. A prediction that comes from this model is that the Unique Stress Constraint is very useful knowledge to have – and we can test if children know it before they can identify words in fluent speech. Because this model was explicitly defined, Gambell and Yang were able to manipulate the learning procedure very precisely and make informative predictions about strategies children might use to solve the word segmentation task.

2.3. Grammatical Categorization

Modeling can be applied to 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*). Mintz (2003) explored one strategy children might use to identify words that behave similarly: *frequent frames*.

Frequent frames consist of framing words that cooccur very frequently in the child's experience. For example, in *she eats it*, the frame is *she___it* and the framed word is *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 together. Notably, they don't rely at all 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 (MacWhinney 2000). The modeling results showed that a frequent frame learner can indeed successfully identify words that behaved similarly solely on the basis of their common frames. These categories mapped very well to the "true" grammatical categories like noun and verb. 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 precisely this sensitivity to frequent frames in infants – and yielded promising results.

2.4. Morphology Acquisition

Modeling has been applied to learning morphology - a common problem is the English past tense. The problem itself is one of mapping: given a verb (*blink, sing, think*), map this word form to the appropriate past tense of that verb (*blinked, sang, thought*). The input data to models is usually realistic estimates of the verbs children encounter during learning, derived from resources like CHILDES (MacWhinney 2000). The output of the model is compared against what is known from experimental work about how and when children appear to learn certain past tense forms.

The main point of interest in many morphology models is that there is a division between a regular pattern (e.g. *blink-blinked*) and several irregular patterns (e.g. *sing-sang, think-thought*). Experimental work indicates that many 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) is often the output goal for English past tense models.

The learning procedures of these models usually take great pains to be psychologically plausible, 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 many others) and probabilistic rule-learning models (Albright & Hayes 2002, Yang 2002, Yang 2005, among others). All models are incremental, learning from a single data point at a time. When the models are able to produce the correct learning trajectory, it is because of some precise design feature within the model – perhaps the way data is representated (ex: Rumelhart & McCelland 1986) or what causes the child to posit a regular rule pattern (ex: Yang 2005).

The predictions generated from these models pertain to the causal factors of the performance trajectory. For instance, a model by Yang (2005) predicts that the performance trajectory depends in a very precise way on the number of regular and irregular verbs encountered by the child and the order in which these forms 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.

2.5. Learning Referential Elements

Modeling has also been applied to learning problems that deal with referential linguistic elements such as anaphors, pronouns, and other referring expressions. The interesting property of referential items is that they can only be interpreted if the listener knows what they refer to. For example, the word *one* in English can be used in a referential manner (sometimes known as the *anaphoric* use of *one*): "Jack has a red ball. He wants another one." Most adult speakers of English interpret the second sentence to mean "He wants another *red ball*." So, the word *one* refers to the words *red ball*, and the referent of *one* in the world is (presumably) a ball that is red. This means that correct interpretation of *one* relies on identifying the words *one* is synonymous with (*red ball*), which then leads to the object in the world *one* refers to (a ball that is red). The learning problem for the English child is how to interpret anaphoric *one* when it's encountered.

Several learning models have attempted to tackle this problem, using incremental, probabilistic learning algorithms on the data. Regier and Gahl (2004) and Pearl and Lidz (under review) 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 (2007) created a model that learned the words *one* referred to (e.g. *red ball*) separately and prior to learning the object in the world *one* referred to (e.g. a ball that is red). Thus, the predictions from these models are that children are sensitive to specific aspects of the available data when learning this kind of interpretation rule. As before, because the hypothesis space and input to these models were precisely defined, the models could manipulate both of these and see the results on learning.

2.6. Syntactic Acquisition

Modeling is also useful for learning the word order rules that make up the syntactic system of language. One example of a word order rule involves the formation of yes/no questions in English when the subject is complex. For instance, take the following sentence: "The knight who can defeat the dragon will save the princess." The yes/no question equivalent is this: "Will the knight who can defeat the dragon save the princess?" Importantly, the auxiliary verb (*will*, *can*, *do*, *might*, 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 has very few explicit examples of this rule - importantly, few enough that children's early acquisition of it seems surprising if their hypotheses for possible rules are not constrained (Legate & Yang 2002). Reali and Christiansen (2004) questioned whether a probabilistically learning child might nonetheless infer the correct rule from other simpler examples of yes/no question formation that are more abundantly available in the input. They designed a model sensitive to certain simple statistical information (called *bigrams*) that a child 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 data used as input (which was derived from CHILDES), a model tracking bigrams preferred the correct complex yes/no question over an incorrect alternative.

Kam, Stoyneshka, Tornyova, Sakas, and Fodor (2005), however, worried that this model's success was due to particular statistical coincidences in the specific data set used as input, and so would not perform as well on different data sets. When they tried the same learning model on other data sets of child-directed speech, they found this to be so – 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.

2.7. Syntactic Acquisition Over Time

Another study of learning word order rules capitalized on the potential interaction between language learning in individuals and language change in a population over time. Put simply, in some cases, language change is thought to happen because individuals don't quite learn the same linguistic knowledge as their predecessors. The children learn it well enough to communicate effectively with the rest of the population, but there may be small changes in the probability with which they use certain rules. It is this individual "mis-learning" that causes change to the entire population, as the small individual changes compound over time (Lightfoot 1999).

Pearl and Weinberg (2007) designed a model that tracked change in the order of objects and verbs. Historical data showed that the population used Object-Verb order (*She peaches eats* = "She eats peaches") more predominantly before the change and Verb-Object order (*She eats peaches*) more predominantly afterwards. Moreover, the change proceeded at a specific rate. Their population model was comprised of individuals who learned their probability of using Object-Verb/Verb-Object order by listening to other speakers within the population, and children

learned from other speakers only for a few years after they were born. At no other time was change allowed to an individual's linguistic knowledge, so how children learned strongly influenced the rate of change in the population.

The population model allowed different individual learning models to be used, and Pearl and Weinberg discovered that the population only changed its language usage at the correct rate when individuals learned in specific ways. In particular, individuals needed to learn from a subset of the available input, rather than using all the data that was available to them. The prediction generated from this model is that these learning restrictions are not just in place for children who live during language change, but are in fact part of the way children learn this information even when the language is stable.

2.8. Stress System Acquisition

Modeling can also cover the acquisition of other complex generative systems, like metrical phonology. The system of metrical phonology determines where the stress is in words. For instance, the word *emphasis* has stress on the first syllable 'em', and not on the other two: it is pronounced EMphasis. It turns out that this stress pattern is generated by a system that groups syllables into larger units called metrical feet, and languages vary on how they group syllables. The child's task is to unconsciously infer the rules that lead to the stress patterns in the observable data.

A model by Pearl (2008) examined this learning problem for English, which has many exceptions to the general rules. Child-directed English speech from the CHILDES database was used as input, and the measure of correct learning in the model was whether the English rules could be learned from this data. The results showed that children could succeed if they learn only from highly informative data, and ignore ambiguous data. In addition, learning success was only guaranteed so long as the rules were learned in a particular order. A prediction generated from this model is that English children should learn the English rules in that special order if they are in fact using only highly informative data to infer the rules.

2.9. Modeling Study Summary

This section has highlighted different studies where modeling complements more traditional experimental research techniques for learning a variety of linguistic knowledge: sounds, words, grammatical categories, morphology, referential elements, and complex systems that generate word order and stress patterns. In each case, the strength of the model is in its empirical grounding and its ability to make predictions that can lead to further experimental research.

3. Subjects

The question of subjects corresponds in modeling to what kind of subject the model is *of*. In all the modeling studies mentioned in the previous section, the simulated learner was a normal monolingual (L1) speaker learning from monolingual data. However, modeling can certainly be extended beyond the normal L1 learning situation, as long as the appropriate input data is available.

As an example, a second-language (L2) learning model could be set up that learns from L2 data. However, the way to distinguish an L1 model from an L2 model is that the L2 model will likely already have linguistic information in place from its own L1. The way this is instantiated in the model will depend on what is known empirically about how L2 learners represent their L1 language rules. The important thing is to ground the model theoretically and empirically. A theoretical grounding will include a description of the knowledge an L2 learner

has of their L1, how it is represented, and how this representation is altered or augmented by data from the L2 language. An empirical grounding will include the data learners have as input and what information they are likely to use to interpret that input (in the L2 learning case, bias from the L1).

More generally, modeling different kinds of subjects requires a detailed instantiation of the relevant aspects of those subjects (knowledge known, initial bias, and interpretation bias, for instance). If this information is available or can be reasonably estimated, a learning model can easily be designed for that subject.

In a similar way, the age of the simulated learner can vary. It is usually set to be the age when the knowledge in question is thought to be learned – 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 is usually 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).

All kinds of learners can be modeled. The key is the model will only be informative if the relevant information about the subject is represented in the model. So, it is important to consider what the relevant information about the subject actually *is* before designing the model. This relates to the next section where we'll review some practical considerations of model design.

4. Description of procedure

For modeling, the relevant experimental procedure is, of course, the model itself. And it's simply the case that models are often 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 parameters of learning that a theory may be vague about – for example, how much data the child processes before learning the relevant information and how fast the child alters her linguistic knowledge when learning. The not-so-good aspect of this concreteness is that the modeler is forced to make an estimate about a reasonable value for unknown parameters.

4.1. 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 learned. Other times, the modeler will simply have to choose a value for convenience and see if this strongly impacts the results of the model. The learning rate in the model, for example, usually requires a value for specifying how much a single data point impacts the child's current hypotheses.

The point is that these parameters affect the outcome of the learning model. So, the value of these parameters may matter. A good way to check this is to try a range of values for the unknown parameters and see the effect on the learning model. If the model's behavior remains invariant, then these parameters, while necessary for implementing the model, do not really affect learning. 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 learning model. For example, if the learning 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 learning the linguistic knowledge in question.

4.2. Control Conditions and Experimental Conditions

From a certain perspective, models are similar to traditional experimental techniques. Experimental techniques usually 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.

There is another way for models to have a control and test condition that is more transparently related to traditional experimental techniques. This has to do with models that simulate a child's ability to make some kind of generalization. Suppose a model is simulating a child's ability to categorize sounds into phonemes, as in the Vallabha et al. (2007) study. The model first learns from data in the input set – individual sounds from child-directed speech in the sound category study. To gauge the model's ability to generalize correctly, the model must then be tested. The sound category model may be given a new sound as input and then output the category that sounds belongs to. The control condition would give the model sounds that were in its input – that is, sounds the model has already encountered and, in fact, 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 – that is, 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 (the way children presumably do), or if it is simply good at classifying the data it's familiar with.

As we recall, data for models often comes from databases of child-directed speech. So, test condition data may come from a different speaker within that database. If the model has not learned to generalize the way children do, 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 (2004). While it was successful when tested on one set of data, Kam et al. (2005) showed that it failed when tested on another set of data. 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 that data. So, it is wise for a model that uses one of these procedures to show that it performs well on a variety of data sets. This will underscore the model's ability to generalize. Since this is a property human language acquisition has, a model able to generalize will be more informative about the main questions of language acquisition.

4.3. More Practical Details

In general, a model will require a computer capable of running whatever program the model is built in. In some cases, the program will be a package where the modeler can simply change the relevant variables and run it on the computer. An example of this is the PRAAT framework designed by Paul Boersma (Boersma 1999), which allows a modeler to test the learnability of sound systems using a particular learning algorithm.

In general, however, the modeler will need to write the program that implements the necessary learning algorithm and describes the relevant details of the simulated learner. In this case, a working knowledge of a programming language is vital – some useful ones are Perl, Java/C++, and Lisp. Knowing at least some of these will give you great flexibility when modeling. Often, it will not take a large amount of programming to implement the desired model in a particular programming language. The trickier part is often in the design of the model itself.

The modeler must consider what information it is important to represent in the simulated

learner. How does the learner represent the required information (sounds, syllables, words)? Does the learner have access to additional information during learning (meanings of words when learning about their sounds, stress contours of words when learning about their meanings)? How does the learner interpret the available data (does the learner need to separate words into syllables, does the learner already know what grammatical category words are)? How will the learner learn (bigrams, tracking frequencies of certain information)? As mentioned earlier, theories are not usually explicit about all these details – but a model must be. Therefore, the modeler will often spend a good deal of time making decisions about these questions before ever writing a single line of programming code.

4.3 Summary of Modeling Procedures

The most crucial aspect of modeling is the decision process behind its design, not the details of how it's programmed. For this reason, this section has focused on the kinds of decisions that are most relevant for language acquisition models. All these decisions focus on how the model will represent both the learner and the learning process. As theories often do not specify all the details a modeler needs to implement the model, the modeler must draw on other sources of information to make the necessary decisions – experimental data and electronic databases like CHILDES provide some guidance. But, the modeler's ingenuity is required to successfully integrate whatever information is available into the design of the learning model. This is a very real component of using models for language acquisition research.

5. Analysis and outcomes

Modeling results can be presented in numerous ways, depending on what the model is testing. Below, we'll review some common methods of representing modeling results.

5.1. Models That Extract Information

For tasks where the model is extracting 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 understand these two measurements, let's switch momentarily to the task of a search engine like Google. Google's job is to identify web pages of interest when it's given a search term (e.g. "goblins", "1980s fantasy movies", "David Bowie"). An 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 will be low. Conversely, the search engine might return only a single relevant web page. Its precision will then be perfect (all returned pages were relevant), but its recall is very 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.

Let's transfer this to some models we've already discussed. One example is the word segmentation model of Gambell and Yang (2006). 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 Mintz (2003). Given a stream of words, the model clusters words appearing in similar frequent frames. Then, these clusters are compared against real grammatical categories to see how well they match. A cluster is assigned to a grammatical category such as *noun* or *verb*. Precision is calculated by dividing the number of words falling in that grammatical category within the cluster (say, the number of 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 (all the verbs in the cluster) by the total number of that grammatical category in the data set (all the verbs in the corpus). It turns out that precision is nearly perfect, but recall is very low. So, this learning method is very accurate in its classifications, but not very complete in classifying all the words that should be classified a particular way.

5.2. Models That Match Children's Performance

Some models simulate the trajectory of children's performance. So, the results show the model's performance over time. This can then be matched against what is known about children's performance over time. A few examples should help illustrate this method.

Often, models of English past tense acquisition (e.g.Rumelhart & McClelland 1986, Yang 2005, among many others) will try to generate a "U-shaped curve" of performance on verbs, which has been observed in children. Specifically, the model will aim 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"), which is then followed by a period where the performance returns to high. A successful model generates this trajectory without having the trajectory explicitly programmed in. The model then aims to explain children's behavior by whatever factor within the model generated this learning trajectory.

Another example of matching trajectories comes from the language change study of Pearl & Weinberg (2007). In that study, individual learning controlled the linguistic composition of a population of speakers over time. Data available from historical records suggests that the population being modeled changed its linguistic composition at a particular rate. The purpose of the model was to match that rate of change. When the model can match the historical trajectory, then we can again examine what factors caused the model to generate the observed trajectory. By this, we can understand the cause of the actual change observed in the historical population.

5.3. Models That Reach A Certain Knowledge State

One might also choose to measure how often a model succeeds at learning. For instance, in the Vallabha et al. (2007) study, the model's goal was to correctly cluster individual sounds into larger language-specific perceptual categories. Different learning 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. In general, this kind of measurement demonstrates the robustness of the learning method. Ideally, we want a learning method that succeeds all the time, since nearly all children successfully acquire the knowledge necessary to be a native speaker.

5.4. Models That Generalize

A related measurement involves testing how often a model makes a correct generalization after being trained on data that children learn from. In the word order studies of Reali & Christiansen (2004) and Kam et al. (2005), models learned how to form yes/no questions like "Is the king singing?" and "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 learned the correct generalization.

5.5. Results Presentation: Summary

The main point of this section is really to highlight that there are a variety of ways to present modeling results. As might be expected, the most effective measure for a model depends on the nature of the model – that is, on what learning task it is trying to simulate. 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.

6. Advantages and disadvantages

Every model is, of course, different. However, we can still discuss the main advantages and disadvantages of modeling without getting into the nuances of individual models.

Put simply, the main advantage of modeling is the ability to manipulate language learning in a very precise way and see the results of that manipulation on learning. In general, the manipulation should be something difficult to do with traditional experimental techniques. For example, it would be very tricky for a traditional experiment to manipulate the hypotheses children entertain, the interpretive biases they impose on the data, or the update procedure they use to shift belief between competing hypotheses. So, modeling provides an effective way to test learning proposals related to these aspects of the learning mechanism.

The main disadvantage is simply that we can never be absolutely sure that our model is really showing how learning works in children's minds. Perhaps some crucial information has been left out of the simulated child's knowledge. Perhaps some critical oversimplifications have been made about how the simulated child interprets the available data. Perhaps the output of the model doesn't have the nuances that children's behavior does. This is why modelers strive for as much empirical grounding as possible. The more checkpoints on the model, the more we can believe what the model shows us about learning. 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 learning behavior that can then be tested experimentally (for example, Pearl (2008)). In this way, experimental research and modeling research can continue to inform each other.

7. Dos and don'ts

Do:

- Read history.
  Learn from previous models about what's reasonable to use as input, algorithms, and measures of output. Consider the strengths and weaknesses of prior models when designing your own.

- Listen with care to linguists.
  Linguists can provide you with the theoretical basis for your hypothesis space, and offer

empirical data to base your model upon.

- Listen with care to psychologists and computational linguists.
Psychologists will also provide you with empirical data to ground your model. Computational
linguists will provide you with learning methods you can implement within a model, and adapt to be psychologically plausible as necessary.

Don't:

- Model when it is obvious.
Models of obvious questions are not informative.

- Forget to ground your model theoretically and empirically.
Models that don't use available data (both theoretical and experimental) as checkpoints are not as persuasive.

- Overlook that this is a model of *human* language acquisition.
Considerations of psychological plausibility should be taken seriously.

References

Albright, A. & Hayes, B. 2002. "Modeling English Past Tense Intuitions with Minimal Generalization". In Mike Maxwell, (ed), *Proceedings of the 2002 Workshop on Morphological Learning*, *Association of Computational Linguistics*. Philadelphia: Association for Computational Linguistics.

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.

Foraker, S., Regier, T., Khetarpal, A., Perfors, A., and Tenenbaum, J. 2007. Indirect evidence and the poverty of the stimulus: The case of anaphoric *one*. In *Proceedings of the 2007 Cognitive Science conference*, Cognitive Science Society.

Gambell, T. & Yang, C. 2006. Word Segmentation: Quick but not dirty. Ms. Yale University.

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. 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.

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.* Boston: Cascadilla Press.

Pearl, L. & Lidz, J. Under review. When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. Ms. 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. 2004. Structure Dependence in Language Acquisition: Uncovering the Statistical Richness of the Stimulus. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society,* Cognitive Science Society.

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*. Cambridge, MA: MIT 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.

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.