Computational Models of Syntactic Acquisition

Charles Yang
Dept. of Linguistics & Computer Science
University of Pennsylvania
charles.yang@ling.upenn.edu

June 2010

Abstract

The computational approach to syntactic acquisition can be pursued fruitfully by integrating results and perspectives from computer science, linguistics, and developmental psychology. In this review, we first examine the statistical properties of natural language that highlight the challenges facing the language learner, which are further elucidated by the mathematical study of learning from several frameworks. We then review two major approaches in computational modeling, one focusing on the exploitation of distributional information in the input while the other relies on a constrained hypothesis space of grammars. By concluding with a discussion of how the computational approach may connect with the empirical study of child language, we make the case for computationally tractable, psychologically plausible and developmentally realistic models of acquisition.

1 Introduction

All models strive to represent reality, and efforts in language research are no exception. Computational models of language acquisition must begin and end as an integral part of the empirical study of child language.\(^3\) A theory of child language cannot be regarded as complete unless it articulates, with sufficient detail, the mechanism responsible for the changes in the linguistic system during acquisition. In other words, while it is useful to demonstrate that “the child knows A at age X but B at age X+Y”, a fuller explanation will require a specification of what kind of learning model, acting on what kind of linguistic data, facilitate the change from A to B during the time course of Y. This is where computational models of learning, which demand a concrete algorithmic process that interacts with the input data in specific ways, can make greatest contribution.

At the same time, it is also critical that computational models be constantly guided and constrained by the findings from the linguistic and psychology studies of child language. It is true that all models make simplifications and idealizations; it is also true that our understanding of language and cognition mind is subject to revision and change. But this uncertainty of knowledge should not give the computational researcher the license of “anything goes” that plagued the early days of Artificial Intelligence. Models of language learning are only informative when formulated with a clear understanding of the linguistic properties that are being learned, the psychological mechanisms of learning that a child is (and is not) capable of using, and the developmental patterns of child language that the model is designed to replicate and explain. Moreover, the search for an acquisition theory applicable across languages should likewise be reflected in the computational approach, which must address the apparent diversity and complexity of the world’s languages.
These themes will be developed throughout this review. We start with a statistical look at natural language data, the raw materials for the acquisition process. We then give a brief survey of learnability research, the general study of learning from data, followed by the discussion of specific formulations that include distributional learning of the grammar and the parameter setting approach to child language. Finally, we develop connections between computational models and developmental studies, a direction that deserves fuller attention in future research.

2 Data and Generalization

A hallmark of human language is its unbounded generative capacity. It is evident in children even, and especially, when they commit linguistic mistakes. Every time a child says “Don’t giggle him” or “The sun is sweating me”, there is a grammatical system at work that generalizes beyond a finite sample of input—and they occasionally get it wrong.

It is useful to examine the properties of natural language, the input to language acquisition, and understand the statistical basis for linguistic generativity. Recent developments in the psychological study of language acquisition have made such studies even more pertinent. First, the very notion of linguistic productivity has been challenged by the item/usage-based approach to language learning and similar linguistic theorizing. For instance, the central tenets of Construction Grammar views constructions as “stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns” and “the totality of our knowledge of language is captured by a network of constructions” [Ref 8, p219]. If so, the generalization problem become a moot point, as the child learner only needs to memorize the constructions in the input. Second, there is now considerable interest in the utility of distributional learning to language, and thus a diminished role for domain specific and/or innate constraints has been supposed in the traditional literature: what can be extracted from the data with general learning abilities needn’t be specially designed or built in.

We return to the discussion of distributional learning in section 4. The role of memorization through item/usage based learning and the accumulation of constructions, however, has been greatly exaggerated.

It is well known that human language follows the so-called Zipf’s law: relatively few words are used frequently while most words occur rarely–on a long tail–with many occurring only once if at all. More precisely, let \( f \) be the frequency of the word with the rank of \( r \) in a set of \( N \) words, then:

\[
f = \frac{C}{r}
\]

where \( C \) is some constant

Zipf’s law can be visualized by taking the log on both sides of the equation above (\( \log f = \log C - \log r \)). A perfect Zipfian would be a straight line with the slope -1, which has been strongly and consistently confirmed in vocabulary studies across languages and genres.

The characteristic long tail of Zipfian distribution becomes more pronounced when we consider combinatorial linguistic units. Figure 1 plots the ranks and frequencies of syntactic rules in the form of context free grammars, based on the parsed modern English corpus, the Penn Treebank. The Zipf-like pattern can be seen by the close approximation by a straight line on the log-log scale.
Figure 1. The frequency distribution of the syntactic rules in the Penn Treebank.

In computational linguistics, Zipf’s law has created the sparse data problem, perhaps the greatest challenge to overcome: the number of linguistic combinations, and thus the number of parameters for a statistical model of language, grows a great deal faster than the amount of linguistic data, as Figure 1 illustrates. These observations are not only relevant for the theory of how the child learns—like the computer, the child does not have unlimited data or infinite time—but also for what the child learns.\textsuperscript{12-13} The item-based nature of child language, and its lack of productivity, is largely established on the basis of the lack of diversity in syntactic combinations.\textsuperscript{4,14} But the combinations of linguistic units (rules, phrases, constructions, etc.) in adult languages follow similar Zipf-like patterns. Figure 2 presents the construction frequencies involving the top 15 most frequently used transitive verbs frequencies from 1.1 million child directed English utterances from the CHILDES database.\textsuperscript{15} The frequencies of the top 10 verb-object constructions are tallied; for instance, “tell him” and “tell John” are different constructions following [Ref 4]. Even for these highly frequent verbs, the probability with which they combine with distinct objects follows a Zipf-like pattern.\textsuperscript{16}
The Zipfian distribution of linguistic combinations means that most constructions as “pairings of form and function” [Ref 8] that one readily uses and processes will never be heard–never mind stored–in any realistic sample of language data. Indeed, statistical analysis of the usage distributions of early child syntax reveals no difference from those of a fully productive grammar.\textsuperscript{12-13} Thus, the child–and the computational model of the child–must be able to generalize far beyond the input: the formal treatment of this process is the learnability study, to which we turn presently.

3 Learnability

In a typical setting of formal learnability, the learner is presented with a sequence of examples drawn from an unknown target language, which can be viewed as a set of strings composed of an alphabet. The learner’s task is to learn the target language after only seeing a finite number of examples (since nobody learns forever). Pertinent to our discussion are two related but distinct frameworks of learning, both of which have developed a very large technical literature. The inductive inference framework\textsuperscript{17} generally requires the learner to converge exactly on the target language within a finite amount of time.\textsuperscript{18-20} The Probably Approximately Correct (PAC) framework\textsuperscript{21} only requires the learner to get arbitrarily close, e.g., the “distance” between the conjectured grammar and the actual grammar can be made as small as possible, but it must be able to do so efficiently (in terms of the amount of examples used, for instance).

Both frameworks are broad enough to allow modifications of the assumptions about the learner, the presentation of the data, the criterion for convergence, etc. In general, however, both frameworks have yielded learnability results that are overwhelmingly negative. For instance, Gold shows that when using positive data alone, only languages that have a finite number of sentences are learnable. Natural language, well known for its infinite use of finite means, lies outside of the learnable class. If negative data is allowed, all primitive recursive functions, which include context free and
context sensitive languages in the Chomsky hierarchy, become learnable. Of course, given the general lack and ineffectiveness of negative evidence in language acquisition and the cross-cultural differences in the mode of parent-child interaction, the cognitive relevance of learning with negative data is at best questionable. When computational complexity is taken into account, as in the case of PAC learning, virtually no language family of linguistic interest, e.g., finite state, context free and context sensitive languages, can be learned efficiently, even if the learner has access to both positive and negative data.

The computational learning theory is well understood but its implications for empirical study of language acquisition deserve some careful consideration and have in fact been misunderstood in many circles; see [Ref 20, 25-26] for clear discussion of these matters.

First, learnability results are very general and can be modified to accommodate a wide range of learning situations. For instance, one could modify the notion of language to include not only strings but string-meaning pairs, giving the child access to semantic information; the non-learnability results hold equally for these alternative conceptions of language. Second, learnability results are usually obtained irrespective of the learning algorithm. In other words, barring major surprises in computational complexity theory that presumably impact our lives more than child language acquisition (e.g., the security of widely used encryption schemes), a negative learnability result is negative not because we have not found an algorithm that works but because no such algorithm can exist. And there is no point in trying the latest and trendiest techniques on a more powerful computer.

Positive learnability results can be achieved but only at the cost of additional assumptions that require independent motivation. One way to do so is to restrict the class of languages with special constraints. While a class of all finite state languages are not learnable, a subclass of finite state languages, the reversible languages, is learnable. Informally, a reversible language describes a set of strings that are guaranteed to have substitutable substrings, a property which allows generalization based on a limited sample of data. This result leads to an efficient learning model for the English auxiliary and noun phrase specifier systems, both of which are well known for their complexity. However, natural language syntax is well known to be above the descriptive power of finite state languages, and the learnability of reversible languages thus has limited applicability.

Another way to obtain learnability is to assume that the learner has access to some additional information about the input data. If the learner has access to the surface string of a sentence as well as its underlying deep structure, then a transformational grammar is learnable under certain learning strategies if the transformational component is suitably constrained. Of course, it could be argued that if the learner could somehow gain direct access to deep structure, from which a good deal of semantic information can be retrieved, there is no little point in bothering with learning a grammar at all.

Finally, the statistical distribution of the language could prove useful if it is available to the learner. Both the inductive inference and the PAC frameworks aim to derive learnability results in the “distribution-sense” sense, that is, no prior assumptions about the distribution from which the learning sample is drawn. This requirement produces results of greatest generality and interest but it can be relaxed as well. It has been shown that if one has certain information about the distribution of the input, then the class of learnable languages is considerably enlarged. But this is a very strong assumption to make, as the estimation of the distribution of a function is generally harder than the estimation of the function itself—and it is the function itself that the child learner is trying to learn during the course of language acquisition: the child tries to learn how to say “I am hungry”, not how often “I am hungry” is said.

It appears that the utility of distribution in learnability has been misunderstood in a wide range of literature in computational linguistics, language acquisition, and linguistics, and almost
always in the context of arguing for a probabilistic concept of learning and grammar, or arguing against the necessity of an innately constrained hypothesis space. The source of these claims appears to be a result due to Horning, a probabilistic instantiation of Gold’s learning paradigm in a Bayesian framework. As reviewed above, context free languages are not learnable under the inductive inference or the PAC learning framework. Horning’s result involves probabilistic context free grammars, which associate context free grammar rules with expansion probabilities. In a probabilistic context free grammar, the probability of a sentence is the product of the probabilities of rules that lead to its derivation. It follows, then, that longer sentences are vanishingly unlikely. Horning’s learner can, in effect, ignore sufficiently long sentences without affecting the overall approximation to the target. Now the grammar is, in effect, finite, a position that few language scientists would find appealing. Finite languages, however, are learnable, as Gold had already shown. Furthermore, Horning’s results are achievable only through exceptional computational resources. Foreshadowing recent work on Bayesian models of learning (section 4), Horning’s algorithm works by searching through the space of probabilistic context free grammars. Not only are these grammar be available to the learner, their prior probabilities of these grammars must also be assumed—this learner makes “more” innate assumptions that the standard inductive learner. The learner calculates the posteriori probabilities of grammars given the data and selects the grammar with the highest value. The computational complexities of Bayesian models are prohibitive, as Horning noted himself. So far as we know, Horning’s model has never been implemented and tested on a reasonable sample of natural language data.

It is interesting that computational models of syntactic acquisition have followed similar paths to obtain positive learnability results. On the one hand, linguistic theories have devoted major efforts to characterize linguistic constraints, Universal Grammar (UG), to ensure that the learner only has access to a limited range of possible languages: the principles and parameter approach in syntax[36] is a prime example (section 5). It is at least a theoretical possibility that the constrained hypothesis space, which is necessary for learnability, comes from the conspiring effects of multiple factors such as working memory, processing limitations,[37] learning mechanisms,[38] rather than the domain specific knowledge of UG alone. To substantiate this possibility, however, requires the specification of exactly what these constraints are, which must also be supposed as innate, and demonstrate their effectiveness. Moreover, at least to the present author, the role of UG in language acquisition is more prominent from the empirical study of language development rather than any mathematical result.[39] On the other hand, the learner may indeed be endowed more powerful computational capacity than generally supposed and is thus capable of successfully exploring a less restrictive space of grammars: this approach is best represented by models of distributional learning of language, which we review below.

4 Grammar and Distributional Learning

The recent flurry of interest in the distributional information of language is frequently seen as a reaction to generative grammar, but that seems to be a misreading of history. Distributional and statistical approach to language and language learning have roots in the structuralist tradition of American linguistics[40] and are evident in the earliest writing on generative grammar.[41] Virtually all current methods of distributional learning, including the clustering based approach to syntactic categories, the information theoretic conception of the grammar and related Bayesian learning models,[42] and the use of transitional probability in word segmentation[43-44] were considered at length. More generally, linguistic theorizing always turns on the notion of distribution, starting with the identification of phonemes in a language and ending with the broadest typological generalizations across
languages. It would certainly be interesting if this process, typically carried out by trained professionals, can be operationalized by the child during the course of language acquisition. The central question is whether distributional learning methods provide adequate account for linguistic structures for the learner: if so, what kind of linguistic principles do they replace, and if not, what kind of linguistic principles ought to be amended.

4.1 Distribution and syntactic categories

A major research area in the distributional learning of language has focused on the acquisition of linguistic categories such as phonemes, words, grammatical categories. These efforts are usually those of data exploration, rather than a psychological model of learning. The following quite is representative: a model of distributional learning via cluster analysis “is not to model the actual procedure a child might use, but rather to examine the information available in children’s input.”

Perhaps the most comprehensive study of distributional learning of syntactic categories is the work of Redington et al., who present a clustering analysis for syntactic categories that formalizes earlier proposals. A word in the child-directed speech is represented by a context vector, which represents the identities of two immediately adjacent words to its left and right; presumably, category membership can be identified by similar usage profiles. Traditional considerations have identified problems with this approach. “John drinks coffee every day” and “John drinks heavily every day” may incorrectly group “coffee” and “heavily” into the same category, and there are linguistic subtleties such as “John is easy/eager to please” that do not straightforwardly fall out of distributional analysis at a surface level. It is thus hoped that spurious generalizations of this sort can be avoided by cluster analysis if the amount of input data is sufficiently large, such that informative data would overwhelm the misleading kind.

Redington et al. uses a hierarchical agglomerative clustering algorithm applied to the vector representations of words. It attempts to merge sufficiently similar word groups, as measured by vector distance, into classes of increasing sizes and thereby creating a tree-like structure of categories. This algorithm is computationally expensive, as the distance between words and groups need to be calculated in pairwise and iterative fashion. Moreover, a decision must be made as to when to stop merging, for otherwise every word will be placed into a same category eventually. The threshold for “cut” is the most critical feature of this type of clustering algorithms; in Redington et al.’s study, a value is hand selected to maximize performance. Free parameters like this often feature in the distributional analysis of language; for instance, other models of syntactic category learning require the researcher to specify the number of categories. These interventions on the part of the researcher raise challenging questions about the psychological plausibility of distributional learning, as it is not obvious that the child finds the optimal parameter values. And there is no clear evidence that favors any specific clustering algorithm.

Distributional information has been shown to be informative in the learning of syntactic categories, but the problem is still far from being resolved. For instance, even the best clusters produced by the Redington et al. model scattered proper nouns across several categories and at the same time grouped wh-words and auxiliary verbs together. And these results are only obtained for the most frequent 1000 words in their corpus of child-directed speech. Given the Zipfian distribution of natural language (section 2), the remaining words are likely to be very infrequent, and there may not be sufficient data to generate their distributional vectors for clusterings. The syntactic category learning problem has been studied in computational linguistics literature as the task of part-of-speech tagging. Considerable progress has been made in supervised learning, where the model has access to a corpus where all words have been manually assigned with the correct category labels. When such data is unavailable, as is the case of child language acquisition, progress has been slower and
the hand tuning of parameters to optimize performance is still the norm.\textsuperscript{50-52}

### 4.2 Distributional learning of grammar

Compared to category learning, there is a comparative dearth of work on distributional learning of grammar, with much of the effort devoted to the acquisition of auxiliary inversion in English question formation, a problem which featured prominently in the argument from the poverty of stimulus to motivate the innateness of the Principle of Structure Dependence in syntax.\textsuperscript{53-55}

Simple recurrent networks\textsuperscript{56} have been trained to discriminate grammatical strings that follow the inversion rule and those that do not (e.g., moving the first auxiliary verb such as “Is the man that tall is nice?”). However, the training data for the network is generated by a very small artificial grammar that consists of only short declarative sentences containing an auxiliary verb and an inverted A\textsuperscript{n}-gram based approach\textsuperscript{57} has been used to capture the patterns of auxiliary inversion based on naturalistic data from child-directed speech. But the effectiveness of this work is simply due to the fact that bigrams such as “who is” are much more frequent than “who tall” in input data, an accidental property of the English relative clauses.\textsuperscript{58} Indeed, the model performs very poorly for other cases of inversion and other languages.

A Bayesian approach represents another recent effort to reduce Structure Dependence to distributional information.\textsuperscript{59} Strictly speaking, the model here does not actually learn a grammar: it evaluates and selects between two grammars, a finite state grammar and a context free grammar, both of which are manually constructed by the researchers. The Bayesian model is similar to Horn-ing’s original approach;\textsuperscript{35}: the grammars are assigned prior probabilities, which are then multiplied with the likelihood of the training data given the grammars, and the grammar with the higher posterior probability is selected—which is the context free grammar. While these researchers state explicitly that they “do not focus on the question of whether the learner can successfully search space” and only study an ideal learner [Ref 59], theoretical considerations\textsuperscript{60} and simulation results suggest that the enormous computational demand on the Bayesian learner may even limit its utility in practice. For instance, just one part of the Bayesian learning model took 352 hours, or 11 days, for a simplified subset of over child-directed utterances (about 15,000 in total).\textsuperscript{61} The type of computing requirement, which grows exponentially when scaled up to realistic samples of linguistic input, is likely to prove taxing for the model’s applicability as well as the modeler’s patience.

A distinct, and potentially fruitful, line of research in the distributional learning of grammar is more directly motivated by experimental findings; see [Ref 39] for a study of statistical and structural learning strategies in word segmentation. This approach may do well to draw insights from the computational linguistics literature where the distributional learning of grammar has been extensively studied as a statistical parsing problem. For instance, a direction in the experimental approach to distributional learning makes use of transitional probabilities between words/categories;\textsuperscript{62-63} adjacent units that are reliably predicted are assumed to constitute part of a syntactic rule, much like the treatment of word segmentation over syllable sequences.\textsuperscript{43} This approach is in fact subsumed by earlier statistical parsing work.\textsuperscript{64} Computational linguists have long noted\textsuperscript{65} that grouping units with high transitional probabilities very often produces incorrect grammatical rules. For instance, nouns and prepositions are frequently and incorrectly grouped into a phrase since English nouns and prepositions are frequently adjacent, which is only because a noun phrase is frequently adjoined by prepositional phrase. Over the past twenty years, various remedies have been proposed to address this problem, largely by introducing more linguistically motivated structures to constrain grammar induction.\textsuperscript{66-68} It would be interesting to pursue similar lines in the experimental approach to see if human subjects can exploit these structural constraints in conjunction with distributional learning.
5 Learning as Selection

The syntactic theory of parameters is usually associated with the Government and Binding theory and the subsequent development of Minimalism. Learning considerations in this framework, which is the focus of this section, extends to any theory of grammar that acknowledges the finiteness of human language grammar: the task of acquisition is to select the grammar(s) used in the learner's linguistic environment from this finite albeit potentially very large set of grammars. Even learning models that use context free grammars may be construed as an instance of parameter setting: the learner is to determine the forms of expansion rules (and their probabilities in a stochastic formalism), assuming, as it is usually the case, that there is an upper bound on the number of non-terminal and terminal nodes, the length and format (e.g., Chomsky Normal Form) of rules. The Bayesian learning model reviewed earlier is similar, as the learner is to choose between a finite state and a context free grammar. In all these approaches, the hypothesis space may be broadly viewed as Universal Grammar (UG), presumably innately available to the learner: the issue is not about the innateness of UG but its particular formulation (e.g., as syntactic parameters or context free grammar rules), as we turn to some helpful evidence from child language in section 6.

5.1 Parameter Setting

The learner’s task, often referred to as parameter setting, is to determine the values of the parameters in her language. The conception of parameters draws firstly from the cross-linguistic comparative work. They can be viewed as a type of anchor points for dividing up the linguistic space: the interactions among them would provide coverage for a vast array of linguistic data—more “facts” captured than the number of parameters, so to speak—such that the determination of the parameter values would amount to a simplification of the learning task. The idea of triggering could be related to the notion of imprinting in ethology: the learner is innately primed to rapidly adopt specific behavioral patterns in the environment.

An influential algorithmic formulation of triggering is given below:

(2) At any time, the learner is identified with a grammar $G$, i.e., a string of 0’s and 1’s
   a. Upon receiving an input sentence $s$, analyze (e.g., parse) $s$ with $G$
   b. If successful then do nothing; return to a.
   c. If failure then
      i. Randomly select a parameter value and flips its value, thus obtaining a new grammar $G’$
      ii. Analyze $s$ with $G’$
      iii. If successful, then adopt $G’$; return to a.
      iv. If failure, revert back to $G$; return to a.

Unfortunately, the triggering model is known to have serious defects. We turn to the developmental issues in section 6; the computational problems of the triggering model have been insightfully analyzed using Markov Chains, a suitable framework for studying all learning models that traverse through a finite space of hypothesis. It turns out that even small parameter spaces have proved problematic for the triggering model. The most serious problem comes from ambiguity between data and grammar. For instance, consider a child learning English, an SVO grammar, but his current hypothesis is a Japanese-like SOV grammar. Suppose the input sentence is “John likes Bill”, for which the SOV grammar fails. There are multiple ways of modifying the grammar that will succeed. For example, the learner could flip the ordering of OV to VO to obtain the target. But it could also
turn on the verb second parameter, which is characteristic of many Germanic languages, in effecting getting the German-like grammar where the underlying word order is SOV (like Japanese) but the movement of the verb to the second clausal position, also leading to the successful parsing of “John likes Bill”. There are no reliable methods to guide the learner toward the target grammar.\textsuperscript{76-77}

There have been two main lines of attacks on the ambiguity problem in parameter setting. The first focuses on how the learner may make more intelligent choices in the navigation of the parameter space. Building on the similar problem of learning the metrical stress system,\textsuperscript{78} it has been proposed\textsuperscript{79-80} that the learner is innately endowed with crucial piece of linguistic patterns dubbed cues, which can reliably determine the values of parameters they are associated with. Thus parameter setting is essentially pre-programmed: the child simply follows the path by looking for specific patterns in the linguistic data. A very similar proposal, the idea of parameter hierarchy,\textsuperscript{73} largely motivated from a comparative/typological considerations, may similarly benefit the child’s task for parameter setting. The natural question, of course, is to what extent the parameters required to describe the words languages follow the ideal expressed in these works.

A related proposal for disentangling parametric ambiguity is suggested by Fodor and colleagues.\textsuperscript{77, 81-82} It avoids learning from input that is compatible with multiple hypotheses and only modifies the grammar on unambiguous data; the detection of ambiguity is to allow the learner try out multiple grammars on an input sentence. Feasibility is a concern, however, as one cannot realistically expect the learner to try out all, or even very many, grammars for any given sentence.

An altogether different approach introduces a domain general and probabilistic learning component to language acquisition under Universal Grammar. The variational learning model\textsuperscript{3} treats the learner as a population of grammars whose probabilistic distribution changes in response to the input, following a process first studied in the mathematical psychology literature.\textsuperscript{83} Suppose that there are \(n\) (binary) parameters \(\alpha_1, \alpha_2, \ldots, \alpha_n\), each parameter \(\alpha_i\) is associated with probability \(p_i\), which denotes the probability of \(\alpha_i\) set to, say, the value 1. The learner is then identified with a \(n\)-dimensional vector of real numbers in \([0, 1]\) \(P = (p_1, p_2, \ldots, p_n)\), and it is \(P\) that changes during the course of learning. A specific instantiation of the variational learning model is illustrated below:

\begin{enumerate}
  \item Upon receiving an input sentence \(s\), the learner uses \(P\) to probabilistically (and thus non-deterministically) generate a composite grammar \(G\).
  \item If \(G\) can analyze \(s\), reward all the parameter choices in \(G\); i.e., increase/decrease \(p_i\) if \(\alpha_i\) has been chosen the value 1/0
  \item If \(G\) fails to analyze \(s\), punish all the parameter choices in \(G\)
\end{enumerate}

Many variants of (3) are possible. For instance, the learner may only reward success but never punish failures. The probability update function can be linear as in Bush & Mosteller’s original formulation or it could be sigmoid or Hebbian. A problem with the model is that for any given trial, a correct parameter value may be punished, and an incorrect parameter value may be rewarded, because the composite grammar \(G\) fails/succeeds: the commitment to online learning does not endow the learner with computational resources to identify exactly the parameters responsible for parsing failure or success. Nevertheless, this naive type of parameter learning provably converge,\textsuperscript{84} but convergence time can be unfeasibly long.

To fully assess the plausibility of parameter setting models would require a more realistic assessment of just what the actual space of human language grammar is. We review some recent efforts in below.
5.2 Toward feasible parameter setting

As we have emphasized earlier, computational modeling is the abstract study of language acquisition. Computational results, both positive and negative, can only be interpreted in the context of empirical findings. For instance, assuming the finiteness of human language grammars as suggested by most linguistic theories, a learner can simply list all grammars in some systematic fashion and sequentially process input sentences from a target grammar. If a grammar is contradicted by a sentence, the learner will move on to the next grammar on the list. It is easy to see that the learner will eventually find the target grammar—all non-target grammars will eventually be rejected—and permanently stay there. Yet no one has seriously proposed this as a plausible model of human language acquisition. It was observed long ago that a major requirement for a theory of grammar is the feasibility of language acquisition: “(w)e want the hypotheses compatible with fixed data to be 'scattered' in value, so that choice among them can be made relatively easily. This requirement of feasibility is the major empirical constraint on a theory, once the conditions of descriptive and explanatory adequacy are met” [Ref 1, 61-62].

Parameters, which can be viewed as a low dimensionality description of syntactic complexity, hold special promise. While the linguistic input consists of abundance of ambiguity with respect to grammars, it is possible that the ambiguity problem is less severe for parameters. The cue-based learning approach, and the notion of parameter hierarchy allow the learner to focus on one parameter at a time and the resolution of each parameter value effectively cuts the space of grammars in half. The probabilistic nature of the variational learning can take advantage of the parametric space in a different way. Specifically, many parameters may be associated with signatures. The signature for a parameter refers to sentences that are analyzable only if that parameter takes on the correct value of the target language. For instance, consider the verb to tense raising parameter, which places the finite verb before negation and adverbs in languages such as French and after in languages such as English:

\[
\begin{align*}
\text{(4)} & \quad \text{a. Jean voit souvent Marie.} \\
& \quad \text{Jean sees often Marie.} \\
& \quad \text{b. John often sees Marie.}
\end{align*}
\]

The relative position of finite verbs and adverbs, then, would be the signature for the verb-to-tense raising parameter: when a child learning French has the parameter to the English option, it is guaranteed to fail upon seeing sentences such as (4a), whereas the English learner cannot analyze (4b) if it has selected the French option either.

In section 6 we outline several more parameters and their signatures, which have important implications for the quantitative study of language development. If all parameters have signatures (with non-trivial frequencies), then the variational learning model, specifically one which only rewards success, can efficiently set parameters. If a signature for a parameter appears in the input, and if the parameter has been chosen for the wrong value, the composite grammar described in (3) must fail, and nothing happens. However, if the parameter has been chosen for the target value, then the composite grammar has some positive probability of succeeding. (Positive, but not absolutely, since the target value of the parameter is a necessary but not sufficient condition for successful analysis of signature.) Thus, when the signature of a parameter is presented in the input, the parameter has a positive probability of moving toward the target and will eventually converge.

There is now reason to be optimistic. Recent work has carried out an extensive exploration of a large and linguistically realistic parameter space (and thousands of languages) and examined its structural properties with respect to learnability. Almost all of the parameters turn out to have signatures, or what these researchers call global triggers. A small minority of parameters have
effective signatures given some other parameters, who do have signatures, have already converged to the target values. For the reward only variational learning, this will merely delay but will not affect convergence.

To summarize, the effort in the study of parameter setting has consistently pointed to the need for a hypothesis space favorable to the learner, which echos the general conclusion from the mathematical study of learning. Assuming the hypothesis space is favorable—given [Ref 82] and the fact that children do learn grammars impressively fast and accurately—there may exist a range of computational learning models that are formally successful. The comparative merits and deficiencies of these models can only be revealed when we turn to the empirical study of child language acquisition.

6 Learnability and Development

It must be said that the connection with language development is the aspect of computational learning that demands most attention and remedy. Some notable early efforts\textsuperscript{2, 85} contain numerous observations about child language and suggestions for the computational mechanisms of language acquisition. The discussion, however, does not generally involve formal treatment though they did led to much subsequent computational work (e.g., distributional learning reviewed earlier). Berwick’s Subset Principle\textsuperscript{86-87} is perhaps the first major result from learnability research to have a direct impact on language acquisition.

6.1 Subset Principle

The Subset Principle follows from the logic of inductive inference:\textsuperscript{17, 19} the hypotheses the learner entertains must be ordered in such a way that positive examples can disconfirm incorrect ones. This tends to force the smallest possible grammar to be adopted first. The Subset Principle can be implemented either as a constraint on the hypothesis space or as a principle of learning that strives for the most conservative generalizations, and these efforts needn’t be mutually exclusive. One of the earliest applications of the Subset Principle concerns the acquisition of grammatical subjects across languages.\textsuperscript{88} The pro-drop grammar such as Italian and topic-drop grammar such as Chinese, which allow the omission but do not prohibit the presence of the subject, appear to constitute a superset to English-like grammar for which the subject is obligatory. The Subset Principle would then imply that the learner adopts the more restrict English option initially. This, however, leads to the prediction that children learning English acquire the obligatory use of subject very early, which would be the default/subset option. Yet the extended period of subject drop, to which we return in a moment, is one of the most important and consistent findings in child language. Upon further reflection, however, it becomes clear that there are no super/subsets among these three types of grammars. The English type obligatory subject grammar is uniquely exemplified by the use of expletive subjects; in the terminology introduced earlier, (5) is the signature for the negative settings of the pro-drop and topic-drop parameters.

(5) a. There is a toy train on the floor.
   b. There seems to some noise in the basement.

It remains to be seen if there are any parameter for which the alternative values constitute a strict subset-superset relation.

A learner that operates by conservative generalizations, which has featured in recent linguistic and psychological theorizing,\textsuperscript{32, 89} can be seen as an embodiment of the Subset Principle as a learning mechanism. A related strategy is the use of indirect negative evidence:\textsuperscript{36} roughly, if the learner had
conjectured an overly general hypothesis but has not observed attestations of examples that would follow that hypothesis, he may be led to retreat to a more restrictive hypothesis. It is possible to develop a Bayesian instantiation of this principle. While the use of Subset Principle may seem appealing, its implementation introduces serious complications. At the very minimum, the determination of superset-subset relations between hypotheses in general involves the calculation of, and subsequent comparisons between, the extensions of grammars (i.e., the set of sentences that they can generate), which can easily become computationally intractable.

6.2 Parameters and Development

Hyams’ pioneering work was the first major effort to make use of parameters and triggering in this regard. It is a well established fact that children acquiring English omit a large amount of subjects, on average 30% of the time, during the first three years of life. A small but non-trivial amount of objects are omitted as well. It is an extremely attractive proposition to assume that children in this stage of acquisition have incorrectly set to the subject parameter to the pro-drop or topic-drop option. Unfortunately, quantitative studies show that the frequencies of subject and object omissions in child English differ significantly from those in (child and adult) Italian and Chinese. Moreover, the disappearance of null subjects is gradual rather than abrupt, as would have been predicted by the triggering model where the child resets an incorrect parameter value. Indeed, the largely gradual development of syntax would pose a challenge for any parameter setting model that makes categorical decisions.

It is of course possible that the mechanisms of grammar learning is not reflected at all in language development. This would be true if by the time when the child produces naturalistic speech, or by the time when experimental techniques can be applied to investigate the nature of child language, all the major aspects of the grammatical system are firmly established with respect to the target language. While this position may disappoint those engaged in the computational modeling of language, developmental researchers from a variety of perspectives have indeed made this assumption though see [Ref 4-5]. There is no denying that many components of child language are learned very early, a pattern that has been repeatedly observed, it is also the case that young children do not talk quite like adults; the null subject phenomenon is just a case in point. Such errors have subsequently been interpreted as deficiencies in either the child’s competence or performance system, both of which may still be in the process of biological maturation that is presumably independent of experience. However, a cross-linguistic look at language acquisition raises serious questions for these approaches. For instance, both Italian and Chinese children, from a very early stage, use subjects and objects at frequencies comparable to adults, in sharp contrast to children learning English. It seems awkward to suggest that obligatory subject languages delays cognitive development and maturation.

The variational learning model was designed to connect the gap between language learnability and development. The introduction of probabilistic learning is designed on the one hand to capture the gradualness of syntactic development and on the other to preserve the utility of parameters in the explanation of non-target forms in child language, all the while providing a quantitative role for the input data in the explanation of child language. Here we briefly summarize the statistical evidence for parameters in syntactic acquisition. For parameters with signatures (see section 5.2), it is clear that those with more frequent signature evidence in the input will be learned—i.e., the probability converging to the target value—faster than those for which signature evidence is less abundant. By estimating the frequency of signatures in child-directed input, one can study the acquisition of parameters quantitatively and cross-linguistically. Table 1 summarizes the results from these investigations; see [Ref 44] for references cited therein for the linguistic and developmental
details, including the empirical evidence for the parameters and their associated signatures.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target</th>
<th>Signature</th>
<th>Input Frequency</th>
<th>Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>wh fronting</td>
<td>English</td>
<td>wh questions</td>
<td>25%</td>
<td>very early</td>
</tr>
<tr>
<td>topic drop</td>
<td>Chinese</td>
<td>null objects</td>
<td>12%</td>
<td>very early</td>
</tr>
<tr>
<td>pro drop</td>
<td>Italian</td>
<td>null subjects in questions</td>
<td>10%</td>
<td>very early</td>
</tr>
<tr>
<td>verb raising</td>
<td>French</td>
<td>verb adverb/pas</td>
<td>7%</td>
<td>very early (1;8)</td>
</tr>
<tr>
<td>obligatory subject</td>
<td>English</td>
<td>expletive subjects</td>
<td>1.2%</td>
<td>3;0</td>
</tr>
<tr>
<td>verb second</td>
<td>German/Dutch</td>
<td>OVS sentences</td>
<td>1.2%</td>
<td>3;0-3;2</td>
</tr>
<tr>
<td>scope marking</td>
<td>English</td>
<td>long-distance questions</td>
<td>0.2%</td>
<td>&gt;4;0</td>
</tr>
</tbody>
</table>

Table 1. Statistical correlates of parameters in the input and output of language acquisition.

Evidence such as this not only provides support for a probabilistic model of learning that is sensitive to the quantity of linguistic input but also highlights the developmental correlates of the parameter based approach. The learning mechanism of the variational model is general and not limited to specific assumptions about the theory of grammar, and is strongly similar to a wide range of machine learning algorithms such as reinforcement learning\(^\text{101}\) that have found applications in many domains and tasks. One can easily apply the learning model to other conceptions of the hypothesis space. If the child’s UG consists of probabilistic context free grammar rules, a small fragment may be:

\[
\begin{align*}
S & \rightarrow^{\alpha} \text{pronoun } VP \\
S & \rightarrow^{\beta} \text{VP, where } \alpha + \beta = 1
\end{align*}
\]

(6) may be viewed as a model of the distribution of pronominal subjects across languages. For English, \(\alpha\) would be close to 1, i.e. all pronoun subjects must be present, whereas in Italian, \(\alpha\) will be fairly small (and \(\beta\) large), i.e., most of pronoun subjects are omitted due to pro-dop. A probabilistic learning model trained on some English and Italian corpora may quickly drive \(\alpha\) and \(\beta\) to the right. But it ought to be obvious that child language development poses some interesting challenges for the surface-based conception of UG such as (6). The overwhelming amount of pronoun subjects in English will push a probabilistic learner very rapidly toward \(\alpha = 1\), but as we have reviewed, the actual learner of English goes through an extended stage of omitted subjects. An explicit learning model, which provides a causal connection from grammar to grammar learning, may play a crucial role in the development and evaluation of linguistic theory.

7 Conclusion

Computational modeling has been an important component of cognitive science since its inception yet it has not been an unqualified success. For instance, computer chess, originally conceived as a case study in human problem solving,\(^\text{102}\) has become essentially an exercise in hardware development, offering no insight on the human mind even as it consistently top-plies the greatest.\(^\text{103}\)

The task of learning a grammar, something that every five year old accomplishes effortlessly, has so far eluded computational brute force. For a research topic that lies at the intersection of linguistics, engineering, and developmental psychology, future research must strive to incorporate the explanatory insights from linguistic theory, to assimilate the formal rigor of computational sciences, and most important, to build connections with the empirical study of child language.
References


