Computational Models of Syntactic Acquisition

Charles Yang
Dept. of Linguistics & Computer Science
Institute for Research in Cognitive Science
University of Pennsylvania
charles.yang@ling.upenn.edu

April 2011

Abstract

The computational approach to syntactic acquisition can be fruitfully pursued by integrating results and perspectives from computer science, linguistics, and developmental psychology. In the article, we first review some key results in computational learning theory that have immediate consequences for language acquisition. We then turn to examine specific learning models, some of which exploit distributional information in the input while others rely on a constrained space of hypotheses, even as both approaches share a common set of characteristics to overcome the learning problem. By concluding with a discussion of how computational models may connect with the empirical study of child grammar, we make the case for computationally tractable, psychologically plausible and developmentally realistic models of acquisition.

1 Introduction

All models strive to represent reality, and the computational study of grammar learning should likewise form an integral part of the empirical work. Language acquisition research typically focuses on the nature of the child's linguistic knowledge–"the child knows A at age X but B at age X+Y"–but a more complete explanation will require a specification of what kind of learning mechanism, acting on what kind of linguistic data, can facilitate the transition from A to B during the time course of X and Y. This is where computational learning models, which demand a concrete algorithmic process that interacts with the input data in specific ways, can make important contributions.

It is equally important that computational models be guided and constrained by the findings from the linguistic and psychological studies of child language.[1–3] The uncertainty in our knowledge about human computational capacities should not issue a blanket license of anything goes. Furthermore, the learning model must produce behavioral patterns consistent with the longitudinal development of grammar which has been amply documented. The search for an acquisition theory applicable across languages should also be reflected in computational studies, which must address the diversity and complexity of the world's languages.

We will develop these themes throughout this article. Section 2 reviews some key results from computational learning theory and highlights the necessity of constraints on the learner that are assumed, in one form or another, by all acquisition models. Section 3 discusses the role of distributional learning in grammar and underscores its connection with computational linguistics where similar topics have been studied. Section 4 focuses on models of acquisition that can be broadly framed as a problem of selecting a target among a finite range of options, with special attention to the complexity and psychological plausibility of the models. Section 5 discusses the need for computational models of grammar acquisition to address the empirical findings in child language development.

2 Learnability

A hallmark of human language is its unbounded generative capacity. This is evident in child language acquisition even–and especially–when children 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 the input, even though it occasionally gets it wrong.

Learnability is the mathematical study of language learning from examples. It is part of the field of computational learning theory which was initiated in part to model child language acquisition.[4–6] Computational learning theory was developed in parallel with the research in statistical inference and approximation[7] and some points of contact between these two traditions can be found.[8] In learnability studies, one typically partitions the problem of learning into several components concerning the presentation of data, the composition of the hypothesis space, the mechanism and complexity of the learning algorithm, the condition of convergence, etc. These components can be varied, producing different learning scenarios that can be studied formally.

Pertinent to our discussion are two related but distinct frameworks of learning. Gold's classic inductive inference framework[4] typically requires the learner to converge exactly on the target language within a finite amount of time and on all the orders in which the examples are pre-

sented. The Probably Approximately Correct (PAC) framework[6] 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. Both frameworks are broad enough to allow variant instantiations of the learning model. In general, however, the theoretical results from both frameworks have been overwhelmingly negative. For instance, Gold shows that when using positive data alone, only the class of finite languages is learnable; none of the classes of languages in the Chomsky hierarchy (regular, context free, context sensitive, recursively enumerable) is learnable. These classes are also unlearnable in the PAC learning framework, which requires computational efficiency, even if the learner has access to both positive and negative data.[9] In fact, even the class of finite languages ceases to be PAC-learnable.[11]

Computational learning theory is well established but its implications for language acquisition require further elucidation; see references (10–11) for clear reviews with special reference to language. First, learnability results are very general and can be modified to accommodate a wide range of learning situations. For instance, the input may consist of form-meaning pairs, e.g., a string and its associated semantics, rather than just the string itself as has been conventionally assumed. The language to be identified would then be a subset of the universe that is the product of the set of all possible strings and the set of all possible meanings: non-learnability results still hold. Second, learnability results are usually obtained irrespective of the specific learning algorithm as long as some widely adopted conjectures about computational complexity hold. There is no point employing the latest and trendiest computational techniques to overcome negative theoretical results.

So far so discouraging, but human children do learn languages. Positive results are possible by providing the learner with additional information about the grammars to be acquired, and/or with more powerful or informative ways of processing the learning data. Not all such modifications are reasonable in the context of language acquisition. For instance, an early result shows that negative data enlarge the learnable class of languages in the inductive inference framework[4] (but not necessarily in the PAC framework). A further result shows that if the learner can present queries to an oracle and obtain certain information about the target language (e.g. whether a string is in the language), then the class of learnable language can be considerably enlarged along with efficient learning algorithms.[20] While results of this type have been appropriately influential in the general study of learning and inference (e.g., pattern classification), it is well known that negative data are not necessary for child language learning and the requirement of an oracle for the child learner to consult also appears suspect. We thus limit our review to some results that rely on at least potentially justifiable assumptions.

An important way to gain learnability is to restrict the space of possible languages. Two ma-

jor directions can be identified: they differ in their methodological orientation and are often viewed as divergent but are in fact similar in spirit. An empirical approach is taken in modern linguistic theorizing, which is devoted to providing a sufficiently restrictive syntactic system for cross linguistic descriptions.[13] To the extent that these efforts are successful, one can take up the question whether they provide plausible computational models of learning; we turn to these issues in section 4 and 5. A more computational approach aims to define demonstrably learnable classes of languages. The central challenge is then to show that such classes are sufficient for the description of human language syntax.

For example, while the entire class of regular language is not learnable, a subset of regular languages with special properties is. An important positive result was given by Angluin.[14] A *reversible* language is a subclass of finite state languages where if two strings share any "tail" (a substring that continues to the end), then they also share *all* tails. For instance, suppose a reversible language contains "John likes pizza", "Mary likes pizza", and "John drinks tea". Since "John" and "Mary" share the same tail ("likes pizza"), they must share all continuations. Thus, "Mary drinks tea" must also be part of the language: the learner thus generalizes. (Sub)string substitutability is the defining characteristic of reversible languages and captures certain intuition about distributional learning.[15] Reversible language induction has been used to learn fairly complex aspects of natural language syntax.[16] However, the class of finite state languages, which properly include reversible languages, is well known to be inadequate for the description of human language syntax. Also, the learnability of reversible languages and similar results for restrictive classes of languages do not do away with the so-called innateness assumption. The learner must "know" that the relevant domain of language is reversible; only then is the deployment of the learning algorithm warranted.

Positive learnability results can also be obtained by providing the learner with additional information about the input. Specific models of grammars are learnable if the learner can access certain structural information about the input string in addition to the string itself. For instance, Wexler & Culicover[17] show that the *Aspects*-style transformational grammar[1] is learnable under certain additional assumptions if the learner has access the D-structure of the sentence, which in effect limits the totality of transformational operations the learner needs to consider. Similar results have been obtained for certain types of categorical grammar[18] and Minimalist grammar[19].

A third way to obtain positive results is to modify/loosen the condition on learnability. The inductive inference and the PAC frameworks, and the research in computational learning theory in general, aim to derive learnability results in a "distribution free" sense, that is, the learner needs to succeed without prior knowledge about the distribution from which the learning sample is drawn. However, if one has certain information about the source distribution of each lan-

guage in the target set, the class of learnable languages is considerably enlarged (though the source distribution itself may be difficult to learn).[10, 20–21] A well known but often misunderstood special case concerns the Bayesian learning approach to probabilistic context free grammars (PCFG).[22] Under a probabilistic context free grammar, longer sentences are exponentially less likely as their probabilities are the product of probabilities of rules used in their derivations. Informally, the learner can ignore sufficiently long sentences without affecting the overall approximation to the target; positive learnability thus can be obtained on a finite (albeit very large) language.[11] Additionally, learnability in this case is achieved by enumerating and evaluating the entire space of possible grammars. These operations are computationally prohibitive even if one ignores the psychological requirements of language learning it is not clear whether this and similar results[23–24] are plausible models of syntactic acquisitions.

3 Grammar and Distributional Learning

The recent flurry of interest in the distributional and statistical information of language is frequently seen as a reaction to generative grammar, but that seems to be a misreading of history. The distributional approach to language and language learning have roots in American structuralist linguistics.[25] It is also evident in the founding documents of generative grammar,[26] which explicitly advocate distributional and information-theoretic approaches to linguistic categories, grammar, and the degree of acceptability etc., as seen in current research.[27–29] Indeed, distributional information is what guides linguists in the structural analysis of languages; it would be of great interest if this process, typically carried out by trained professionals, can be operationalized by the child during the course of language acquisition.

Much of the statistical parsing research in computational linguistics can be viewed as applications of distributional learning to grammar. This line of work typically differs from the goals and methods of language acquisition: unlike the child, a statistical parser is "supervised" as it has access to a parsed corpus, and there is no need to justify the psychological validity of the learning algorithm. The state of the art statistical parsers[30–31] can produce useful results but there is still much room for improvement in both quality and efficiency that would approach human level analysis. The present article is not the appropriate forum to review that vast literature; instead, we will touch upon some insights from statistical parsing in connection to current theorizing in linguistics and psychology.

There is comparably little work on unsupervised learning from text, a task closer to that of language acquisition, and none has produced results approaching the quality of supervised statistical parsing. Much effort has been devoted to a subproblem in grammar acquisition, the

auxiliary inversion rule in English questions, which has featured prominently as to demonstrate the principle of structure dependence in syntax.[32–33] One set of results is discriminative in nature: a distributional learning model is trained to distinguish grammatical examples of auxiliary inversion from ungrammatical ones (e.g., moving the first auxiliary verb such as "Is the boy that _ tall is nice?"). A simple recurrent network[34] can be trained for this purpose. However, the training data for the network are generated by a very small artificial grammar and it is not known how the model would fare in face of realistic child-directed data. Simple statistical models of language such as n-grams also seem to recognize the correct pattern of auxiliary inversion.[35] Subsequent study[36] shows that this result is due to the fact that bigrams such as "who is", which appears in the grammatical string "Is the boy who is tall _ nice" are much more frequent than "who tall", which appears in the ungrammatical string "Is the boy who _ tall is nice", a reflection of the numerous short Wh-questions in child-directed English (e.g., "who is here?"). The n-gram model performs very poorly for other cases of inversion and for languages such as Dutch where question formation does not have the (accidental) property of English that works in favor of the model.

Bayesian learning models, which have gained popularity in cognitive science, have also see applications to the problem of auxiliary inversion.[23] Strictly speaking, the Bayesian model does not actually learn a grammar: it evaluates and selects one out of two types of grammars, a finite state grammar and a context free grammar, both of which are manually constructed by the researchers from a simplified subset of child-directed English. The selection of the target among a pool of candidates is Bayesian while other criterions such as the Minimum Description Length principle may also be used.[37] (In this sense, the Bayesian model is more in line with the parameter setting approach to language acquisition (section 4 and 5) where learning is viewed as selecting a hypothesis out of an innately specific set.) Like Horning's formulation of Bayesian learning of grammars,[22] the two grammars are assigned prior probabilities, with the smaller grammar being favored. The learning model then calculates the likelihood of the input data given a grammar, which is then multiplied with the prior probability of the grammar to obtain the posterior probability of the grammar. The model is able to favor the context free grammar when the input data has reached a certain level of volume and complexity. It should be noted that the context free grammar in the Bayesian model already contains rules for the structure dependent inversion of the auxiliary; its relevance to the innateness debate is a moot point. While Bayesian models typically deal with an optimal learner[24] and are often explicit in denying psychological plausibility, theoretical considerations[38] and simulation results[39] suggest that the enormous computational demand on the Bayesian learner may even limit its utility as an idealized model.

A distinct, and potentially fruitful, line of distributional learning research is more directly

rooted by human learning abilities demonstrated in the laboratory. Computational models can help evaluate their effectiveness in a realistic setting,[40] as we review two main results from computational linguistics that are of direct relevance to empirical research. First, recent studies of artificial language learning suggest that syntactic rules might be learned via the use of transitional probabilities between words/categories.[41–42] This approach has been studied in statistical parsing,[43–44] often producing linguistically incorrect rules. For instance, a verb and a preposition are frequently adjacent and may thus be grouped together as a rule but that is merely a reflection of the rule that places a verb immediately before a prepositional *phrase*. The progress in statistical parsing can be attributed to more linguistically motivated structures to constrain grammar induction;[30] it would be interesting to see if these structural constraints can be exploited by human subjects in an experimental setting.

Second, a statistical parser may provide insights on the power as well as limitations of distributional information. For instance, a statistical model of syntax can make use of a wide range of grammatical rules: an phrase "drink water" may be represented in multiple forms ranging from categorical (VP $\rightarrow$ V NP) to lexically specific: (VP $\rightarrow$ V_{drink} NP, or even VP $\rightarrow$ V_{drink} NP_{water}.[45]) In practice, it has been found that lexicalization provides very little gain over simpler models that only use general rules,[46–47] These findings are a reflection of the sparse data problem in computational linguistics,[48] which inherently limits storage-based approaches to learning and lexicalized approaches to grammar. The fundamental problem of language learning, distributional or otherwise, remains to be that of generalization from a small set of data.

4 Learning as Selection

The syntactic theory of parameters is usually associated with the Government and Binding theory and the subsequent development of Minimalism.[13] Formal considerations of learning, however, can be extended to any language model that accepts the finiteness of human grammars. Acquisition in this setting amounts to selecting the grammar(s) used in the learner's linguistic environment from a pre-defined set. Even learning models that use context free grammars, or the Bayesian learning model reviewed earlier, can be viewed as an instance of parameter setting: the learner is to determine the forms of expansion rules (and their probabilities in a stochastic formalism), In all these approaches, the constitutive primitives of the grammar space, which can be broadly called Universal Grammar (UG), are assumed to be innately available to the learner. The occasionally heated debate in language acquisition is not about the innateness of UG but about particular conceptions of UG: e.g., whether the learner should be characterized as a set of abstract parameters or context free grammar rules. The debate is an empirical one and we expect

the evidence from child language to play a role (section 5). For the purpose of the present review, we focus on computational models of grammar selection more directly situated in the Principles & Parameters framework, chiefly due to the amount of empirical child language research in this tradition.

The original motivation for parameters comes from comparative syntax. Parameters may provide a more compact description of grammatical facts than construction specific rules; parameterization of syntax can be likened to the problem of dimension reduction in the familiar practice of principal component analysis. For language acquisition, the learner needs to determine the parameter values for her language. Consider an influential algorithmic formulation known as triggering.[49] At any time the learner is identified with a single parameter setting. The learner randomly changes a parameter value if the current setting fails to analyze an input string. The revised setting is adopted if it succeeds; otherwise the learner reverts back to the old setting before moving on to the next string. The triggering model operates in an online fashion so as to reduce the cognitive load of the learner, and the use of error driven learning follows a long tradition in learnability research.[4, 17, 50] Further analysis of the triggering model,[51] however, reveals serious convergence difficulties. At the heart of the matter is the ambiguity problem between data and grammar. In an error-driven learning scheme, the failure on an input sentence may result in multiple ways of updating the current parameter setting, but there is no reliable way for the online learner to know which ones lead to the target and which ones drift further and further away.

One way to resolve the ambiguity problem is to endow the learner with special knowledge of the parameter domain.[52] In some approaches, parameter setting follows a pre-determined (i.e. innately specified) sequence: the determination of a parameter value before the setting of another may eliminate or reduce the ambiguity problem, and similar ideas have been applied to other parametric domains of language such as metrical stress. A related proposal is to provide the learner with the ability to detect grammar-data ambiguity.[53] The learner may carry out multiple parses for an input string: if more than one parameter settings are successful, then the string is clearly ambiguous and the learner will move on to the next string without altering the current parameter setting. Furthermore, a structural description of the input string may provide additional cues to guide the learner's actions than a simple success-failure check, as has been established in the learnability studies.[17, 18–19]

A different approach introduces a probabilistic, and possibly domain general, learning component to parameter setting. In the variational model,[3] the learner is identified not with a single parameter setting but with a population of parameter settings whose probabilistic distribution changes in response to the input. The mechanism of learning has roots in mathematical

psychology[54] and modern theories of machine learning.[55] A binary parameter α_i is associated with a probability p_i, which denotes the probability that α_i is set to 1. Upon receiving an input string, the learner generates a composite grammar G based on the p_i's. If G succeeds, all the chosen values of the parameters are rewarded; no action is taken if G fails. It is possible that a wrong parameter value may be rewarded if G succeeds thanks to other, correctly set, parameters. For instance, consider a parameter that only concerns interrogative sentences: even a wrong value of this parameter does not affect the analysis of a declarative sentence in the input, for which the parameter is not even relevant and may be incorrectly rewarded. This difficulty does not pose formal barriers for convergence though the worse case complexity is exponential.[56] (Of course, if the learner were to know which parameters are relevant for grammatical analysis, the problem may go away altogether though currently there is no successful proposal on decoding string-parameter relations.) Efficient learning is possible if most parameters have independent "signature" strings for which successful analysis necessarily requires the correct values of these parameters regardless of others.[3]

Little work so far–in either distributional learning and parameter setting–has studied a grammar domain sufficiently complex for cross-linguistic variation; some recent work has given reasons for optimism. Taking 13 linguistically important parameters pertaining to word order variations in the world's languages, Sakas and Fodor have constructed a set of over 3000 "languages" and almost 50,000 distinct syntactic patterns are generated.[57] While the data-grammar ambiguity is high as long expected, the data-*parameter* ambiguity is promisingly low: 10 out of the 13 parameters have independent signatures referred to above,[3] and the remaining three effectively have signatures after the other parameters are set. The space of parameters thus appears favorable to the learner. If so, a wide range of computational learning models may prove sufficient in the selection of the target grammar. The comparative merits and deficiencies of these models can only be revealed when we turn to the empirical study of child language acquisition.

5 Learnability and Development

In most general terms, computational models of syntactic acquisition attempt to find the best combination of grammar models and learning algorithms to account for the developmental findings in child language. Aside from a few notable early efforts, the connection with empirical child language research is an area in computational learning that demands most attention and remedy. Pinker's important contribution[2] contains many suggestions for the computational mechanisms of language acquisition though virtually no formal treatment is given. The Subset Principle[50] is perhaps the first major result from learnability research to have a direct impact on

language acquisition.

Berwick's Subset Principle follows from the logic of inductive inference and is implicit in earlier results:[4, 5] 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: no other grammar compatible with the data that leads to the new grammar should be a (proper) subset of that grammar. 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 and their parametric treatment. The *pro-drop* grammar such as Italian and *topic-drop* grammar such as Chinese, which allow the omission (though 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 imply that the learner adopt the more restrict English option initially. Unfortunately this leads to the prediction that children learning English acquire the obligatory use of subject initially, as it is the subset default option–contrary to the well attested subject drop stage in child English to be discussed below. It turns out that the English grammar is not a subset of the pro-drop or topic-drop grammar: obligatory subject languages such as English are exemplified by the use of expletive subjects (e.g, "there is a car coming") which are not present in pro/topic-drop grammars. 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 both linguistic and psychological theorizing,[58–59] can be seen as an embodiment of the Subset Principle as a learning mechanism. A related strategy is the use of indirect negative evidence:[13] if the learner had conjectured an overly general hypothesis but has not observed attestations of examples that would follow that hypothesis, it may retreat to a more restrictive hypothesis. In other words, absence of evidence *is* evidence of absence: a logically flawed but possibly human principle of inference. The use of indirect negative evidence may be implemented in various ways[60] though there may be serious complications in the execution. At the very minimum, the determination of superset-subset relations involves comparison of extensions of grammars, which appears computationally intractable when we deal with realistically complex grammars.[61]

The theory of parameters offers promise for the empirical study of language development. Since the totality of grammar is capped, the child's systematic errors can be interpreted as biologically possible though non-target grammars. The well known phenomenon of subject drop in child language is a case in point. English learning children omit up to 30% of grammatical subjects during the first three years of life; a smaller but non-trivial number of obligatory objects are

omitted as well. An attractive position is to attribute these errors to a mis-set parameter to the pro-drop (as in Italian) or topic-drop (as in Mandarin Chinese) option though these predictions are not borne out empirically.[62–63] Of course, it remains possible that the children has in fact learned the English grammar correctly very early[2, 62] and the omitted subjects and objects are due to non-syntactic factors such as performance. But cross-linguistic studies reveal difficulties with this approach. For instance, both Italian and Chinese children from a very early stage use subjects and objects at frequencies comparable to adults,[62–63], in sharp contrast to the delay in child English.

The variational learning model may help close the gap between language learnability and language development.[3] 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. And it must be acknowledged that language acquisition research in the generative tradition has not sufficient attention to the role of the input. Here we briefly summarize some quantitative evidence for parameters in syntactic acquisition. Parameters with a larger amount of signatures (section 4) in the input, which can be estimated from child-directed speech data, can be expected to be set faster than those for which signatures are less abundant. It thus accounts for, among other findings, why English children approach the adult use of subjects and objects with an extended delay–as the learner still probabilistically drops the topic–while Italian and Chinese learning children are on target early.

TABLE I HERE

While formal studies of acquisition have received sufficient attention through mathematical and computational analysis, the developmental patterns of child language may provide decisive in the consideration of alternative approaches. Consider the child's hypothesis space (or UG) as a class of probabilistic context free rules. For instance, the rule "S$\xrightarrow{\alpha}$ pronoun VP" may correspond to the requirement of a subject in English, and "S$\xrightarrow{\beta}$ VP" accounts for the fact that languages like Italian allow subject drop: the learner's task is to determine the weights (α and β) of these rules. A probabilistic learning model applied to English and Italian corpora may quickly drive α and β to the right values: $\beta \approx 0$ in the case of English. But one immediately sees that this learning trajectory of PCFG is inconsistent with child language, as English learning children go through an extended stage of subject drop despite the overwhelming amount of overt subjects in the adults' speech. The formal study of syntactic acquisition allows for the manipulation of

the hypothesis space and the learning algorithm to explore their empirical consequences.

6 Conclusion

Computational modeling has been an important component of cognitive science since its inception yet it has not been an unqualified success.Computer chess, originally conceived as a showcase for human problem solving,[64] has become an exercise in hardware development, offering no insight on the mind even as it consistently topples the greatest.[65]

The task of learning a grammar, something that every five year old accomplishes with ease, has so far eluded computational brute force. For a research topic that lies at the intersection of linguistics, engineering, and developmental psychology, progress can only be made if we 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

1. Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
2. Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
3. Yang, C. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.
4. Gold, M. (1967). Language identification in the limit. *Information and Control*, 10:447-74.
5. Angluin, D. (1980). Inductive inference of formal language from positive data. *Information and Control*, 45, 2, 117-135.
6. Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27, 1134-1142.
7. Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.
8. Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 929-965.
9. Kearns, M. & Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM*, 41, 67-95.

10. Osherson, D., Stob, M., & Weinstein, S. (1986). *Systems that learn*. Cambridge, MA: MIT Press.

11. Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

12. . Queries and concept learning. *Machine Learning*, 2, 319-342.

13. Chomsky, N. (1981). *Lectures on government and binding*. Dordrectht: Foris.

14. Angluin, D. (1982). Inference of reversible languages. *Journal of the ACM*, 29, 3, 741-765.

15. Clark, A. & Eyraud, R. (2007). Polynomial identification in the limit of substitutable context free languages. *Journal of Machine Learning Research*, 8, 1725-1745.

16. Berwick, R. & Pilato, S. (1987). Learning syntax by automata induction. *Machine learning*, 2, 1, 9-38.

17. Wexler, K. & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

18. Kanazawa, M. (1998). Learnable classes of categorical grammars. CLSI: Stanford University.

19. Stabler, E. (1998). Acquiring languages with movement. *Syntax*, 1, 72-97.

20. Angluin, D. (1988). Identifying languages from stochastic examples. Technical Report 614. Yale University. New Haven, CT.

21. Pitt, L. (1989). Probabilistic inductive inference. *Journal of the ACM*. 36, 383-433.

22. Horning, J. (1969). A study of grammatical inference. Doctoral dissertation. Department of Computer Science. Stanford University. Stanford, CA.

23. Perfors, A., Tenenbaum, J. & Regier, T. (2006). Poverty of the stimulus? A rational approach. *Proceedings of the 28th annual conference of the Cognitive Science Society*. Vancouver, Canada.

24. Charter, N. & Vitányi, P. (2007). "Ideal learning" of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51, 135-163.

25. Harris, Z. (1951). *Methods in structural linguistics*. Chicago: Chicago University Press.

26. Chomsky, N. (1955/1975) The logical structure of linguistic theory. Manuscript, Harvard/MIT. Published in 1975 by New York: Plenum.

27. Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 4, 425-469.

28. Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358, 1239-1253.

29. Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 153–198.

30. Collins, M. (1999). Head-driven statistical models for natural language processing. Ph.D. dissertation. University of Pennsylvania.

31. Charniak, E. (2000). A maximum-entropy-inspired parser. Proceedings of NAACL, 1, 132-139.

32. Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

33. Legate, J. A. & Yang, C. (2002). Empirical reassessments of poverty stimulus arguments. *Linguistic Review*, 19, 151-162.

34. Lewis, J. & Elman, J. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual Boston University conference on language development*. Somerville, MA: Cascadilla. 359-370.

35. Reali, F. & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29, 1007–1028.

36. Kam, X., Stoyneshka, I., Tornyova, L., Fodor, J. D. & Sakas, W. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*, 32, 771-787.

37. . Hsu, A. & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, 34, 972-1016.

38. Hackerman, D., Geiger, D. & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20, 197-243.

39. McClelland, J. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1, 11-38.

40. Yang, C. (2004). Universal grammar, statistics, or both. *Trends in Cognitive Sciences*, 451-456.

41. Saffran, J. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44, 493-515.

42. Thompson, S. & Newport, E. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3, 1, 1-42.

43. Magerman, D. & Marcus, M. (1990). Parsing a natural language using mutual information statistics. *Proceedings of the AAAI*. 984-989.

44. de Marcken, C. (1995). On the unsupervised induction of phrase-structure grammar. In *Proceedings of the Third Workshop on Very Large Corpora.*. 14-26.

45. Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Cambridge, MA: Harvard University Press.

46. Bikel, D. (2004) Intricacies of Collins's parsing model. *Computational Linguistics*, 30, 479–511.

47. Klein, D. & Manning, C. (2003). Accurate unlexicalized parsing. *Proceedings of the ACL*.

48. Jelinek, F. (1999) *Statistical methods for speech recognition.* Cambridge, MA: MIT Press.

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

50. Berwick, R. (1985). *The acquisition of syntactic knowledge.* Cambridge, MA: MIT Press.

51. Berwick, R., & Niyogi, P. (1996). Learning from triggers. *Linguistic Inquiry*, 27:605-622.

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

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

54. Bush, R., & Mosteller, F. (1951). A Mathematical Model for Simple Learning. *Psychological Review*, 68:313-323.

55. Sutton, R. & Barto, A. (1998). *Reinforcement learning.* Cambridge, MA: MIT Press.

56. Straus, K. (2008). Validations of a probabilistic model of language acquisition. Ph.D. dissertation. Department of Mathematics, Northeastern University.

57. Sakas, W. & Fodor, J. D. (in press). Disambiguating syntactic triggers. *Language Acquisition*.

58. Culicover, P. (1999). *Syntactic nuts*. New York: Oxford University Press.

59. MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31, 883–914.

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

61. Fodor, J. D. & Sakas, W. (2005). The subset principle in syntax. *Journal of Linguistics*, 41, 513-569.

62. Valian, V. (1991). Syntactic Subjects in the Early Speech of American and Italian Children. *Cognition*, 40:21-82.

63. Wang, Q., Lillo-Martin, D., Best, C., & Levitt, A. (1992). Null Subject vs. Null Object: Some Evidence from the Acquisition of Chinese and English. *Language Acquisition*, 2:221-54.

64. Newell, A., Simon, H. & Shaw, D. (1958). Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development*, 2, 4, 320-335.

65. Kasparov, G. (2010). The Chess Master and the Computer. *New York Review of Books*, 57, 2.

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

Table 1: Statistical correlates of parameters in the input and output of language acquisition. Very early acquisition refers to cases where children rarely, if ever, deviate from target form, which can typically be observed as soon as they enter into multiple word stage of production. The 90% criterion of usage in obligator context is used to mark successful acquisition. The references to the linguistic and developmental details of these case studies can be found in (3).