Advanced Review

Computational models of syntactic acquisition

Charles Yang*

The computational approach to syntactic acquisition can be fruitfully pursued by integrating results and perspectives from computer science, linguistics, and developmental psychology. In this article, we first review some key results in computational learning theory and their implications 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, yet both approaches share a common set of characteristics to overcome the learning problem. We conclude with a discussion of how computational models connects with the empirical study of child grammar, making the case for computationally tractable, psychologically plausible and developmentally realistic models of acquisition. © 2011 John Wiley & Sons, Ltd.

How to cite this article:
WIREs Cogn Sci 2012, 3:205–213. doi: 10.1002/wcs.1154

INTRODUCTION

All models strive to represent reality, and the computational study of grammar learning likewise forms an integral part of child language research. Language acquisition studies typically focus 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 from X to $X + Y$. This is where computational learning models, which demand concrete algorithmic processes that interact 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] Uncertainty in our knowledge about human computational capacities should not warrant a blanket license of 'anything goes', eschewing formal and empirical considerations. Furthermore, the learning model must yield behavioral patterns consistent with the longitudinal development of grammar that has been amply documented. Finally, the search for an acquisition theory applicable across languages should also be reflected in computational studies, which must address the world's linguistic diversity and complexity.

We will develop these themes throughout this article. Section on *Learnability* 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 on *Grammar and Distributional Learning* discusses the role of distributional learning in syntactic acquisition and underscores its connection with computational linguistics where similar topics have been studied. Section on *Learning as Selection* 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 complexity and psychological plausibility. Section on *Learnability and Development* addresses the need for computational models of grammar acquisition to connect with the empirical research in language development.

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

*Correspondence to: charles.yang@babel.ling.upenn.edu
Department of Linguistics, Computer Science & Psychology, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, USA

giggle him' or 'The sun is sweating me', there is a grammatical system at work that generalizes beyond the input, and it occasionally gets it wrong.

Learnability is the mathematical study of language learning, which is viewed as the discovery of any computable function from examples to grammars/languages. It is a subarea of computational learning theory which was initiated in part to model child language acquisition.[4–6] Computational learning theory was developed in parallel with statistical inference and approximation[7] and some points of contact between the two traditions can be found.[8] In learnability studies, the learning problem is partitioned 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 which can then be formally studied.

There are two major related but distinct frameworks for learnability study. Gold's 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 presented. The probably approximately correct (PAC) framework[6] only requires the learner to get arbitrarily close (hence 'approximately') to the target with high probabilities (hence 'probably'), but it must be able to do so efficiently. Both frameworks are broad enough to allow variant instantiations of learning model. In general, however, the theoretical results from both frameworks have been overwhelmingly negative. For instance, Gold shows that with positive data, only the class of finite languages is learnable; none of the classes in the Chomsky hierarchy (regular, context free, context sensitive, recursively enumerable) is learnable. These classes are also unlearnable in the PAC learning, which allows negative data but demands computational efficiency.[9]

Computational learning theory is well established but its implications for language acquisition require further elucidation; see Refs 10 and 11 for reviews. 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, for example, 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: Gold's nonlearnability results still hold. Second, learnability results are usually obtained irrespective of the specific learning algorithm if one assumes some widely accepted conjectures about computational complexity: there is no point employing the latest and trendiest computational techniques to overcome negative theoretical results.

Yet children do learn languages. Positive results are possible by providing the learner with additional information about the grammars to be acquired and/or informative ways of processing the learning data. Immediately after proving his negative results, Gold suggests three directions in which learnability can be achieved.[4] First, while the classes in the Chomsky hierarchy are not learnable, it is possible that the class of natural languages is much smaller or more restricted, which limits the learner's choices further. Second, if the child receives negative evidence, then all recursively enumerable languages are learnable in the inductive inference framework (but not necessarily in the PAC framework). Third, the class of learnable languages is also enlarged if there is some priori restriction on the type of data that can occur. For instance, if the learner knows that the nonoccurrence of a string implies its ungrammaticality, that would consitute negative evidence and ensure learnability.

Most subsequent positive learnability results fall into these three categories, but not all results are appropriate for language acquisition. For instance, negative evidence is not systematically available to the learner and is in any case unnecessary for the success of language acquisition. Similarly, having an oracle, to which the learner can present queries about the language (e.g., whether a string is grammatical), yields a larger learnable class along with efficient learning strategies.[12] But query based learning is even more suspect in the context of language acquisition, though it is clearly useful in the general study of learning and inference such as pattern classification. In this review, we provide a brief survey of results that are at least potentially empirically relevant.

An important way to gain learnability is to restrict the space of possible languages—Gold's first suggestion. Two major directions can be identified: they differ in 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 sections on *Learning as Selection* and *Learnability and Development*. A computational approach tries to define demonstrably learnable classes of formal languages. The central challenge is then to show that such formal classes are sufficient for the description of natural language.

For example, while the entire class of regular languages is not learnable, a subset of regular languages with special properties is, following an important positive result given by Angluin.[14] A *reversible* language is a subclass of regular 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'. Because 'John' and 'Mary' share a tail (i.e., 'likes pizza'), they must share all continuations by definition. Thus, if the learner knows that the language in question is reversible, then it can conclude that 'Mary drinks tea' must also be part of the language, thereby achieving generalization. Reversible languages capture the notion of distributions in linguistics,[15] and have been applied to learn fairly complex grammatical aspects of English.[16] But the usefulness of these results is limited because the class of regular languages, which properly include reversible languages, is well known to be inadequate for the description of human language syntax.[17]

In addition to negative evidence, other sources of information about the input may also benefit the learner. Specific models of grammars are learnable if the learner can access certain structural information about the input string in addition to the string itself. An important result is due to Wexler and Culicover.[18] An *Aspects*-style transformational grammar[1] is learnable if the learner has access to the D-structure of the sentence and only needs to consider a bounded domain of examples (e.g., limited by the depth of embedding). These assumptions limit the totality of transformational interactions and errors that a learner might see, thereby making learnability possible. Similar results have been obtained for certain types of categorial grammar[19] and Minimalist grammars.[20] However, it remains unclear to what extent these structural aspects of the input, beyond the string itself, are systematically available to the child learner.

A third way to obtain positive results is to relax the condition on learnability. The inductive inference and the PAC frameworks, and the research in computational learning theory in general, attempt to derive 'distribution free' learnability results; that is, the learner needs to succeed without prior knowledge about the distribution from which the learning sample is drawn. If the source distribution of each language in the target set is independently available, the class of learnable languages is considerably enlarged, at least formally.[10,21–23] For instance, a well-known special case is Horning's Bayesian approach to learn probabilistic context free grammars (PCFG).[21] Under a PCFG, longer sentences are exponentially less likely since the probability of a sentence is the product of probabilities of rules used in its derivation: the learner can ignore sufficiently long sentence without affecting the overall approximation to the target. This effectively constitutes negative evidence, which follows the third strategy to achieve learnability in Gold's original discussion. (We return to the use of probabilistic data as negative evidence in section on *Learnability and Development*.) Although the specific distributional properties of PCFG do benefit learning, PCFG as a class of formal languages is also inadequate for syntactic descriptions.[24] Indeed, we currently do not know much at all about the source distribution of natural languages, so the stringent requirement of distribution free learning still seems most prudent. Additionally, Horning's result is achieved by the enumeration and evaluation of the entire space of possible grammars, a technique used in similar subsequent developments.[25,26] These techniques are prohibitively expensive, and learnability thus obtained should presently be regarded as existence proofs rather than psychological models of acquisition.

GRAMMAR AND DISTRIBUTIONAL LEARNING

The recent flurry of interest in the distributional learning of language is frequently seen as a reaction to generative grammar, but that misses an important part of history. The distributional approach to language has roots in American structuralist linguistics.[27] It is also evident in the founding documents of generative grammar,[28] which explicitly advocate distributional and information-theoretic approaches to linguistic categories, grammar, and the degree of acceptability, etc.; these methods are now actively pursued in distributional learning research.[29–31] 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. Recent advances in computing technology have now made it possible to assess the role of distributional information in language acquisition.

Statistical parsing research in computational linguistics examplifies the distributional learning at the fullest scale. Most statistical parsers are 'supervised': a probabilistic model of grammar has access to parsed corpora with annotated tree-like structures, which enable the valuation of parameters, and the model is then evaluated on novel sentences for coverage and generalization.[32,33] Unsupervised parsing, where the model learns directly from text, is more similar to language acquisition but parsing quality is still well short

of supervised methods. Neither supervised nor unsupervised parsing are intended to be models of human syntactic learning: they tend to involve iterative optimizations over the entire corpus and the learning algorithms are not subject to the psychological constraints on the child learner. In the present article, we will review some results from statistical parsing that have more direct bearing on current theorizing in linguistics and psychology.

A great deal of distributional learning research has been devoted to a specific case in grammar acquisition, the auxiliary inversion rule in English questions, which has featured prominently in demonstrating the principle of structure dependence in syntax.[34,35] 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 can be trained for this purpose.[36] 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 a realistic linguistic environment. Simple statistical models of language such as *n*-grams also seem to recognize the correct pattern of auxiliary inversion.[37] As pointed out in a subsequent study,[38] this is because 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'. 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.[38]

Bayesian learning models, which have gained popularity in cognitive science, have also been applied to the problem of auxiliary inversion.[25] Strictly speaking, the Bayesian model does not actually learn a grammar: it evaluates and selects one of two types of grammars, a finite state grammar and a context free grammar constructed by the researchers from a subset of child-directed English. (Since the context free grammar already contains the structure dependent rule for auxiliary inversion, the innateness issue is a moot point.) The selection of the target among a pool of candidates is Bayesian though other criteria such as the Minimum Description Length principle may also be used[26,28]; in this sense, the Bayesian model is more in line with the parameter setting approach to language acquisition, reviewed in sections on *Learning as Selection* and *Learnability and Development*, where learning is viewed as selecting a hypothesis out of a predetermined set. As in Horning's formulation,[21] the two grammars are assigned prior probabilities, with the smaller grammar being favored. The model then calculates the likelihood of the input data given a grammar, which is then multiplied with the prior probability to obtain the posterior probability of the grammar. The model favors the context free grammar when the input data has reached a certain level of volume and complexity. While Bayesian and similar models typically deal with an idealized learner with unrestricted computational power and are often explicit in not claiming psychological plausibility, theoretical considerations and simulations suggest that the enormous computational demands on the Bayesian learner may even limit its utility as an abstract model.[39,40]

A distinct and potentially fruitful line of distributional learning research is more directly based on human learning abilities demonstrated in the laboratory. Computational models can help evaluate their effectiveness in a realistic setting[41] as we review two main results from computational linguistics that have direct connections to empirical research. First, recent studies of artificial language learning suggest that syntactic rules might be learned with transitional probabilities across words/categories in a sentence.[42,43] This approach has been studied in statistical parsing,[44,45] 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 the co-occurrence is 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[32,45]; it would be interesting to see if these structural constraints can similarly be exploited by human learners.

Second, a statistical parser, particularly one which is not burdened by constraints that a psychological model must be subject to, may provide insights into the utility of distributional information that is in principle available under most favorable conditions. For instance, a statistical model of grammar can make use of a wide range of grammatical rules: a 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 bilexically specific ($VP \rightarrow V_{drink}\ NP_{water}$). When tested on novel data, it has been found that the majority of generalizing power comes from categorical rules; lexicalization plays an important role in resolving syntactic ambiguities[32] but bilexical rules offer virtually no additional coverage.[46] These findings are a reflection of the sparse data problem in computational linguistics,[47] which inherently limits storage/example-based approaches to learning and lexicalized approaches to grammar.[48] The

fundamental problem of language learning, distributional or otherwise, remains to be that of generalization from a small set of data.

LEARNING AS SELECTION

The syntactic theory of parameters is usually associated with the Principles & Parameters framework and the subsequent development of Minimalism.[13] Formal considerations of learning, however, can be extended to any language model that admits only a finite number of possible grammars. Acquisition in this setting amounts to selecting the grammar(s) used in the learner's linguistic environment from a predefined 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: for example, 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 on *Learnability and Development*). 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. In this sense, the theory of parameters is similar to distributional learning methods such as the minimum description length principle as both follow an information theoretic approach to the grammar with respect to a corpus of data.[28] 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 unique 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 resource requirements of the learner, and the use of error driven learning follows a long tradition in learnability research.[4,18,50] Further analysis of the triggering model,[51] however, reveals that the model frequently fails to converge. 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 predetermined sequence: the resolution of one 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 studied in learnability research.[18–20]

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 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, but formal convergence results have been obtained.[56] Furthermore, 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 to allow a kind of 'scattering' favorable to the learner,[1] despite the enormous space of possible grammars. 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.

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.

The 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 hypotheses. 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 need not 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 *prodrop* 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 should 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. Indeed the English type grammar is not a subset of prodrop 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 is 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, one which was suggested in Gold's original study of learnability.[4] But the use of probabilistic data as a substitute for negative evidence requires at least some additional justification, as grammaticality and probability of a sentence are logically separate issues. And there may be complications in the implementation of indirect negative evidence. The determination of Superset–subset relations involves comparison of extensions of grammars, which appears computationally intractable when we deal with realistically complex grammars.[60]

The theory of parameters offers promise for the empirical study of language development. As the total number of grammars is capped, the child's systematic errors can be interpreted as linguistically possible though nontarget 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 3 years of life; a smaller but nontrivial number of obligatory objects are omitted as well. An attractive position is to attribute these errors to a mis-set parameter of the prodrop (as in Italian) or topic-drop (as in Mandarin Chinese) option though these predictions are not borne out empirically.[61,62] Of course, it remains possible that

TABLE 1 | Statistical Correlates of Parameters in the Input and Output of Language Acquisition

Parameter	Target	Signature	Input Frequency (%)	Acquisition
Wh fronting	English	Wh questions	25	Very early
Topic-drop	Chinese	Null objects	12	Very early
Prodrop	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

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 Ref 3.

the child has in fact learned the English grammar correctly very early[2,61] and the omitted subjects and objects are due to nonsyntactic 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,[61,62] 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 nontarget 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 paid 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 on *Learning as Selection*) 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 1).

While formal models 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'shypothesis space (or UG)as a class of PCFG rules; here we follow an early effort that models a fragment of an English learning child's syntax.[63] 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 enables the learning algorithm to explore their empirical consequences.

CONCLUSION

Computational methods have 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 now 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, assimilate the formal rigor of computational sciences, and most important, build connections with the empirical study of child language.

ACKNOWLEDGMENTS

I would like to thank two anonymous reviewers and Peter Culicover, the associate editor, for their very helpful comments and suggestions.

REFERENCES

1. Chomsky N. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press; 1965.
2. Pinker S. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press; 1984.
3. Yang C. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press; 2002.
4. Gold M. Language identification in the limit. *Inform Control* 1967, 10:447–474.
5. Angluin D. Inductive inference of formal language from positive data. *Inform Control* 1980, 45:117–135.
6. Valiant L. A theory of the learnable. *Commun ACM* 1984, 27:1134–1142.
7. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 2000.
8. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Learnability and the Vapnik-Chervonenkis dimension. *J ACM* 929–965.
9. Kearns M, Valiant L. Cryptographic limitations on learning Boolean formulae and finite automata. *J ACM* 1994, 41:67–95.
10. Osherson D, Stob M, Weinstein S. *Systems that Learn*. Cambridge, MA: MIT Press; 1986.
11. Niyogi P. *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: MIT Press; 2006.
12. Angluin D. Queries and concept learning. *Mach Learn* 1988, 2:319–342.
13. Chomsky N. *Lectures on Government and Binding*. Dordrecht: Foris; 1981.
14. Angluin D. Inference of reversible languages. *J ACM* 1982, 29:741–765.
15. Clark A, Eyraud R. Polynomial identification in the limit of substitutable context free languages. *J Mach Learn Res* 2007, 8:1725–1745.
16. Berwick R, Pilato S. Learning syntax by automata induction. *Machine Learn* 1987, 2:9–38.
17. Chomsky N. *Syntactic Structure*. The Hauge: Mouton & Co; 1957.
18. Wexler K, Culicover P. *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press; 1980.
19. Kanazawa M. *Learnable Classes of Categorical Grammars*. Stanford University: CLSI; 1998.
20. Stabler E. Acquiring languages with movement. *Syntax* 1998, 1:72–97.
21. Horning J. A study of grammatical inference, Doctoral dissertation, Department of Computer Science, Stanford University, Stanford, CA, 1969.
22. Angluin D. *Identifying languages from stochastic examples*. Technical Report 614. New Haven, CT: Yale University; 1988.
23. Pitt L. Probabilistic inductive inference. *J ACM* 1989, 36:383–433.
24. Shieber S. Evidence against context-freeness of natural language. *Ling Phil* 1985, 8:333–343.
25. Perfors A, Tenenbaum J, Regier T. Poverty of the stimulus? A rational approach, In Proceedings of the 28th annual conference of the Cognitive Science Society. Vancouver, Canada, 2006.
26. Chater N, Vitányi P. 'Ideal learning' of natural language: positive results about learning from positive evidence. *J Math Psychol* 2007, 51:135–163.
27. Harris Z. *Methods in Structural Linguistics*. Chicago: Chicago University Press; 1951.
28. Chomsky N. The Logical Structure of Linguistic Theory, Manuscript, Harvard/MIT. Published in 1975 by New York: Plenum; 1955/1975.
29. Redington M, Chater N, Finch S. Distributional information: a powerful cue for acquiring syntactic categories. *Cogn Sci* 1998, 22:425–469.
30. Pereira F. Formal grammar and information theory: together again? *Phil Trans Royal Soc* 2000, 358: 1239–1253.
31. Goldsmith J. Unsupervised learning of the morphology of a natural language. *Comp Ling* 2001, 153–198.
32. Collins M. Head-driven statistical models for natural language processing, Ph.D. dissertation, University of Pennsylvania, 1999.
33. Charniak E. A maximum-entropy-inspired parser. *Proc NAACL* 2000, 1:132–139.
34. Chomsky N. *Reflections on Language*. New York: Pantheon; 1975.
35. Legate JA, Yang C. Empirical reassessments of poverty-stimulus arguments. *Ling Rev* 2002, 19:151–162.
36. Lewis J, Elman J. 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, 2001, 359–370.
37. Reali F, Christiansen MH. Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cogn Sci* 2005, 29:1007–1028.

38. Kam X, Stoyneshka I, Tornyova L, Fodor JD, Sakas W. Bigrams and the richness of the stimulus. *Cogn Sci* 2008, 32:771–787.
39. Hackerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach Learn* 1995, 20:197–243.
40. McClelland J. The place of modeling in cognitive science. *Topics Cogn Sci* 2009, 1:11–38.
41. Yang C. Universal grammar, statistics, or both. *Trends Cogn Sci* 2004, 451–456.
42. Saffran J. The use of predictive dependencies in language learning. *J Memory Lang* 2001, 44:493–515.
43. Thompson S, Newport E. Statistical learning of syntax: The role of transitional probability. *Lang Learn Dev* 2007, 3:1–42.
44. Magerman D, Marcus M. Parsing a natural language using mutual information statistics, In: Proceedings of the AAAI, 1990, 984–989.
45. de Marcken C. On the unsupervised induction of phrase-structure grammar, In *Proceedings of the Third Workshop on Very Large Corpora* 1995, 14–26.
46. Bikel D. Intricacies of Collins's parsing model. *Comp Ling* 2004, 30:479–511.
47. Jelinek F. *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press; 1999.
48. Yang C. A statistical test for grammar, In *Proceedings of the ACL*. Portland, OR, 2011.
49. Gibson E, Wexler K. Triggers. *Ling Inq* 1994, 25:355–407.
50. Berwick R. *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press; 1985.
51. Berwick R, Niyogi P. Learning from triggers. *Ling Inq* 1996, 27:605–622.
52. Dresher E. Charting the learning path: cues to parameter setting. *Ling Inq* 1999, 30:27–67.
53. Fodor JD. Unambiguous triggers. *Ling Inq* 1998, 29:1–36.
54. Bush R, Mosteller F. A mathematical model for simple learning. *Psychol Rev* 1951, 68:313–323.
55. Sutton R, Barto A. *Reinforcement Learning*. Cambridge, MA: MIT Press; 1998.
56. Straus K. Validations of a probabilistic model of language acquisition, Ph.D. dissertation. Department of Mathematics, Northeastern University, 2008.
57. Sakas W, Fodor JD. Disambiguating syntactic triggers. *Lang Acquisit* (in press).
58. Culicover P. *Syntactic Nuts*. New York: Oxford University Press; 1999.
59. MacWhinney B. A multiple process solution to the logical problem of language acquisition. *J Child Lang* 2004, 31:883–914.
60. Fodor JD, Sakas W. The subset principle in syntax. *J Ling* 2005, 41:513–569.
61. Valian V. Syntactic subjects in the early speech of American and Italian children. *Cognition* 1991, 40:21–82.
62. Wang Q, Lillo-Martin D, Best C, Levitt A. Null subject vs. null object: some evidence from the acquisition of Chinese and English. *Lang Acquisit* 1992, 2:221–254.
63. Suppes P. Probabilistic grammars for natural languages. *Synthese* 1970, 22:95–116.
64. Newell A, Simon H, Shaw D. Chess-playing programs and the problem of complexity. *IBM J Res Dev* 1958, 2:320–335.
65. Kasparov G. The chess master and the computer. *New York Rev Books* 2010, 57:2.

FURTHER READING

Berwick R, Pietroski P, Yankama B, Chomsky N. Poverty of the stimulus revisited. *Cogn Sci* 2011, 35:1207–1242.

Bertolo S (ed.). *Language Acquisition and Learnability*. Cambridge: Cambridge University Press; 2011.

Pinker S. Formal models of language learning. *Cognition* 1979, 7:217–283.

