Fodor, J.D. and Sakas, W.G. (2017), in: Ian Roberts (ed.)
The Oxford Handbook of Universal Grammar. Oxford, U.K.

Chapter Eleven

Learnability

Janet Dean Fodor

CUNY Graduate Center

William Gregory Sakas

Hunter College, CUNY

11.1 Introduction

The study of natural language learnability necessarily draws on many different strands of research. The goal – as least as we will construe it here – is to construct a psychological model of the process of language acquisition, which reveals how it is possible for a human infant, in little more than 4 or 5 years, to arrive at a sophisticated command of the complex system that is a human language. Such a project straddles the fields of descriptive and theoretical linguistics, developmental psychology, the psycholinguistics of language comprehension and production, formal learning theory, and computational modeling. Building a model of language acquisition must take into account the cognitive capacities (such as attention, memory, analytic powers) of a pre-school child; the facts of the language that the child is exposed to; the relationship between those facts and the language knowledge the child eventually arrives at; and the kinds of data manipulation processes that might in principle bridge the gap between input and end-state knowledge. While the emphasis in the current state of research is on devising *any* learning system that could achieve what children reliably achieve, any particular model will be the more convincing, the more closely its predictions fit the observed developmental course of language acquisition.

Progress has been made on many of these fronts in the last half-century or so, making this a lively and engrossing topic of research. But it must be acknowledged that despite the best of intentions, and some clever and innovative individual research programs, the contributing disciplines have not yet been able to integrate themselves to create a unitary picture of how children do what they do. Part of the reason for this, as will become clear, is that *within* each sub-discipline there are currently unsettled issues and competing views, so that building the overarching model we seek is like trying to piece together a jigsaw puzzle without knowing which of the puzzle pieces on the table belong to it.

A classical disagreement concerns nature versus nurture. Linguists investigate the extent to which natural languages have properties in common: are there linguistic universals? The answer to this could hint at whether learners have innate knowledge of what a language can be like. That in turn may reflect on whether or not the mental mechanisms for acquiring language resemble those for acquiring knowledge of the world in general. A cross-cutting uncertainty arising within linguistic theory concerns how abstractly the structures of sentences are mentally represented in the adult grammar. The answer to that may have implications for whether learners can gain broad linguistic knowledge from specific observations, or whether they have to build up generalizations gradually over many observations.

11.2 Mathematical theorems

The modern study of language learnability was founded in the 1960s following Chomsky and Miller's remarkable collection of papers on the formal properties of natural language, including mathematical modeling of the relations between properties of languages and properties of the grammars that generate them (Chomsky (1959/1963); Chomsky & Miller (1963); Miller & Chomsky (1963)).[1] The learnability studies that ensued were mathematical also, and hence untrammeled by the demands of psychological plausibility that we impose today. Proofs were based on abstractions far removed from actual human languages (e.g., grammars were typically designated merely as g1, g2, g3 …, and the languages they generated as constituted of sentences s1, s2, s3 …). The groundbreaking work by Gold (1967) presented a general algorithm which, given an input sample from a target language, would hypothesize a grammar for that language. Gold showed that the success or failure of this or any other such algorithm depends not only on the properties of the target grammar itself but also on the properties of the range of alternative grammar hypotheses that might be entertained. (Technically, in that framework, what was learnable was thus not a language, but a class of languages.) This has been taken to underwrite the importance of there being an innate specification of the class of possible grammars for human languages (but see Johnson (2004)).

The most cited theorem resulting from Gold's work, known as 'Gold's Theorem', establishes the non-learnability of every class of languages in the 'Chomsky hierarchy' (see Chomsky (1956)) from positive evidence only, i.e., from exposure to nothing but a sample of the sentences of the target language. This is called *text presentation*; it provides information about sentences that are in a target language, but no information about what is not in the language. This result resonated with linguists and psychologists who had observed the dearth of negative linguistic evidence available to children, providing anecdotes of children's resistance to explicit correction (Brown & Hanlon (1970), McNeill (1966)) and more recent systematic data concerning the non-specificity of adult reformulations (Marcus (1993)). Gold's Theorem thereby fueled the conviction in some (psycho)linguistic circles that there must be innate constraints on the space of potential grammar hypotheses to compensate for the paucity of information in a learner's input. (See discussion of 'the argument from the poverty of the stimulus' in Chapter10.) This body of innately given constraints on the grammar space has come to be known as Universal Grammar (UG).

Another important concept emanating from this research, still highly relevant today, is the problem posed by subset-superset relations between languages, for learners without access to negative evidence. Suppose the set of sentences in one possible language is a proper subset of the set of sentences in another one (e.g., the former is identical to English except that it lacks reduced relative clauses). Then a learner that hypothesizes the more inclusive language, when the target is the less inclusive one, would be capable of generating sentences that are ungrammatical in the target – an error that would be detectable by speakers of the target language, but not by the learner on the basis of the evidence available in positive-only input. It is clear that this does not generally happen, because if superset choices could be freely made, languages would historically grow more and more inclusive, as the output of one speaker's grammar is input to later learners. Gold noted that this kind of 'superset' error would be avoidable if the possible grammar hypotheses were prioritized, with every 'subset grammar' taking precedence over all of its 'superset grammars'.[2]

Imagine, then, that learners have knowledge of this prioritization, and systematically test grammars in order of priority, checking whether or not a given grammar generates the current input sentence. The learner would move on from one grammar to the next only if the former fails to generate a sentence in the input sample. Under this strategy a learner will never pass over the target grammar in favor of one of its supersets. In Gold's proofs, this beneficial partial ordering of grammars was folded into a total ordering, which he called an *enumeration*. If human learners employed such an enumeration, it would also be of benefit in another way: it would automatically register which grammar hypotheses have already been tested and failed on some input sentence, making it possible to avoid unnecessary re-testing. Current models with psychological aspirations mostly do not presuppose a full enumeration, for reasons to be given below. However, a partial ordering (possibly innate) of grammars which respects subset-superset relations remains a central presupposition in theoretical studies of language acquisition, under the heading of the *Subset Principle* (Berwick (1985); Manzini & Wexler (1987); see later discussion in Fodor & Sakas (2005)).

Out of Gold's work grew a whole field of research in mathematical learning theory, extending beyond human language learning. For language learnability, other mathematical studies followed Gold's. Horning (1969) introduced a probabilistic approach and obtained a positive learning result: that the class of probabilistic context-free grammars is learnable. Angluin (1980) identified some non-trivial classes of languages that cut across the Chomsky hierarchy, which are learnable on the basis of the original (non-probabilistic) Gold paradigm. Osherson, Stob and Weinstein (1986), also within the Gold tradition, contributed many additional theorems. Of particular interest to linguists was their proof that if a learner's input sample of sentences from the target language could be 'noisy', containing an unlimited number of intrusions of a finite number of word strings drawn from other languages in the domain, then learnability would be attainable only if the domain of languages were finite. This made a timely connection with the recent shift by Chomsky (1981) from an open-ended class of rule-based grammars for natural languages to a finite class of parametrically defined grammars.

These formal studies of learnability had the great merit of delivering provable theorems, with broad application to all learning situations of whatever type was defined in the premises of the proof. If research had been able to proceed along those lines, it could have led rapidly toward the goal of a single viable learning model for language. But that was not possible. In order for such broad-scope proofs to go through, their premises had to be highly abstract and general, divorced from all the particular considerations that enter into psychological feasibility: the actual size of the set of grammar hypotheses; the extent of learners' memory for individual sentences; the specific degree of ambiguity between the input word strings with respect to the grammars they are compatible with; and so forth. While we may not be able to assess these factors with great precision at present, some 'ballpark' comparisons can be made between formal models and human reality and they suggest that there is a considerable gulf between them.

Subsequent commentaries have pointed out respects in which Gold's framing of the learning process, however rigorous and illuminating it is in formal respects, may not be psychologically plausible as a basis for modeling human language acquisition; see especially Pinker (1979). While innate prioritization of grammar hypotheses could indeed eliminate superset errors, Gold's concept of an enumeration as a total ordering of the possible grammars was regarded as making implausible predictions if construed psycholinguistically. Some languages would be acquired significantly more slowly than others, just because they appear late in the innate enumeration. Regardless of how distinctive its sentences might be, a 'late' grammar could not be adopted until all prior grammars in the enumeration had been tried and falsified.[3] On the other hand, grammar ordering might be regarded as a virtue, related to the linguistic concept of markedness, if the grammars earlier in the enumeration were optimal grammars, conforming most closely to central principles of natural language. But Pinker (1979; 234) judged that even so, some 'good' languages would be intolerably slow to attain: "In general, the problem of learning by enumeration within a reasonable time bound is likely to be intractable." [4]

Perhaps the most telling objection against construing the Gold paradigm as the basis for a psychological model is that the grammars hypothesized by such a learner in response to input might bear no relation at all to the properties of that input. For example, consider Gold's assumption that the learner proceeds along the enumeration until a compatible grammar is found. On the not unreasonable assumption that a human learner could test at most one new grammar against each input sentence (or entire input sample), an encounter with a sentence beyond the scope of the current grammar would cause a shift to the next grammar in the enumeration regardless of whether that grammar was any more compatible with the input than the current one – indeed it might be even less so. In contrast, we expect that when children change their grammar hypotheses in response to a novel input sentence, the properties of that sentence guide them toward a new grammar which could incorporate it. It would be odd indeed (cause for concern!) if a child shifted from one grammar to an arbitrarily different one, just because the former proved inadequate. Pinker made much the same point in advocating "heuristic language learning procedures" which hold promise of learning more rapidly because they "draw…their power from the exploitation of detailed properties of the sample sentences instead of the exhaustive enumeration of a class of grammars"[5] (1979: 235). We return to this notion of 'input-guided' learning in section 6 below.

11.3 Psychological feasibility

As learnability theory aimed to become a contributing partner in the modeling of real-life language learning, it had to become more specific in its assumptions, both about languages and about language learners. In place of broadly applicable theorems about learning in general, particular models had to be developed and evaluated. These expanded ambitions proved difficult to achieve. The first noteworthy study in the new (psycho)linguistic tradition was that of Wexler & Hamburger (1973), Hamburger & Wexler (1973, 1975)), followed up by the heroic work of Wexler & Culicover (1980). The aim was to show that it was possible for a linguistically authentic grammar to be acquired by computational operations of a reasonably modest complexity, on the basis of a language sample consonant with the language that toddlers are exposed to. In other words, a learning model for natural language was to be developed which was psychologically feasible, computationally sound, and in tune with current linguistic theory.

Wexler & Culicover's demonstration of learnability made a very specific assumption about what kind of grammar was acquired: a transformational grammar in the general style of the (Extended) Standard Theory (as in Chomsky (1965, 1973)), in which surface structures of sentences were derived by cyclic application of a rich variety of transformations to deep structures defined by context-free phrase structure rules, and only deep structures were semantically interpreted. Wexler & Culicover (henceforth W&C) assumed that the class of possible deep structures was tightly constrained and universal, and also that a learner could establish the deep structure of any given surface sentence (word string) from its non-verbal context via a unique association between deep structures and meanings. This was a particularly strong assumption, whose role was to create a fixed starting point for the typically complex transformational derivation of a surface sentence. Another crucial assumption concerned how the learner responded to a mismatch between the currently hypothesized grammar and the input. W&C proposed that on encountering an input sentence the learner would run through the transformational derivation from the known deep structure to whatever surface word string resulted from it by applying the currently hypothesized transformational rules. If that word string matched the input, no grammar change would be made; thus this learner was 'error-driven' (see below). If it did not match, the learner would select at random a transformation to delete from the current grammar, regardless of its effect on the derived word string; or else it would add at random to the grammar any one (legitimate) transformational rule that would convert the predicted surface form into the actually observed surface form. As we noted in the case of enumeration-based learning, the resulting grammar might be less adequate than the one it replaced (e.g., loss of a needed transformation). This element of randomness in the revision of wrong grammars should be borne in mind as we proceed to other models that followed.

W&C were able to develop an intricate proof that such a learner, under these conditions, would eventually converge on a descriptively adequate grammar of any target language compatible with their version of the Standard Theory, on the basis of reasonably simple input ('degree-2' sentences, i.e., with a maximum depth of two embedded clauses), though only if certain constraints on transformational derivations were respected. One such constraint, which W&C dubbed the Binary Principle[6], was strongly reminiscent of the Subjacency Condition that Chomsky (1973) had argued for on purely linguistic grounds. Other constraints that were required in order for the learnability proof to go through also had some theoretical linguistic credentials. For example the Freezing Principle[7] was shown to cover a number of observed limits on transformational applications. The fact that independently motivated constraints on deep structures and on derivational operations were shown to be crucial contributions to learnability was greeted as further evidence for substantial innate pre-programming of humans for language acquisition.

W&C's work was a milestone in the search for a fully specified learning model that respected both linguistic and psychological considerations. It is true that in this early work some of the assumptions that bracketed the performance of the model were not the most stringent approximations to what might be assumed about child learners and their input (e.g. the necessity of some degree-2 input sentences). But for all that, this study clearly established a new research project for the future: proving natural language learnability within increasingly tighter and more realistic bounds of psycholinguistic feasibility.

11.4 Changing grammars: The move to parameters

No sooner had W&C completed their learnability proof than the linguistic theory which it presupposed (Chomsky's Standard Theory, subsequently Extended Standard Theory) was abandoned. In 1981Chomsky outlined a major revision of the theory of transformational grammar which had consequences both for the formal properties of the grammars that learners were assumed to acquire, and for the mechanisms by which those grammars could be attained. In the course of introducing this new Government-Binding (GB) theory, Chomsky first proposed the concept of *parameters* as the means to codify cross-language differences in syntactic structure. This theory is therefore often referred to as the Principles and Parameters (P&P) theory, and this terminology (though loose) is useful since the concept of parameters has been long retained through various changes in the definition of government and other details of GB grammars, and even into the substantial theoretical revisions resulting in the Minimalist Program (Chomsky (1995b) and since). Chomsky's motivations for this major change in 1981 included both linguistic (typological) matters and learning-related issues. He wrote: "The theory of UG must meet two obvious conditions. On the one hand, it must be compatible with the diversity of existing (indeed, possible) grammars. At the same time, UG must be sufficiently constrained and restrictive in the options it permits so as to account for the fact that each of these grammars develops in the mind on the basis of quite limited evidence. … What we expect to find, then, is a highly structured theory of UG based on a number of fundamental principles that sharply restrict the class of attainable grammars and narrowly constrain their form, but with parameters that have to be fixed by experience" (Chomsky (1981: 3-4).

P&P theory was actually the culmination of a growing trend throughout the 1970s to streamline grammars by eliminating the constraints previously included as part of each transformational rule, in favor of very general constraints on all rule applications (of the sort that figured in W&C's learnability proof). Individual transformations were eventually replaced entirely by the assumption that any possible transformational operation is free to apply at any point in a derivation *except* where specifically blocked by such constraints. The one maximally general transformational rule that remained was known most famously as 'Move α' (Chomsky (1980)) and in even broader form as 'Affect α' (Lasnik & Saito (1992)). This development left no way to register variation between languages in terms of their different sets of transformational rules, but the newly introduced parameters took on this role. All aspects of the syntactic component were assumed to be innate and universal, except for a finite number of choice points offered by the parameters.[8] For example, one basic parameter establishes whether syntactic phrases are head-initial (e.g., the verb precedes its object in a verb phrase) or head-final (the verb follows the object); another parameter controls whether a Wh-phrase is moved to the top (the Complementizer projection) of a clause or is left in situ (see Chapter 14).

In order to have full command of the syntactic structure of the target language, a learner would need only to establish the values of the parameters relevant to it. Thus, the research focus now shifted from how a learner could formulate a correct set of transformational rules, to how a learner could identify the correct parameter settings on the basis of an input sample.

11.5 Implementing parameter setting: domain search

Chomsky himself was not very explicit about the mechanism for setting the syntactic parameters, but he drew attention to important aspects of it. Given a finite number of parameters, each with a finite number of values, the acquisition process would no longer be open-ended, but would be tightly focused. As Chomsky described it (1986b: 146) “We may think of UG as an intricately structured system, but one that is only partially ‘wired up.’ The system is associated with a finite set of switches, each of which has a finite number of positions (perhaps two). Experience is required to set the switches.” The learner would have to spot just one relevant fact about the language for each parameter value, or perhaps just one fact for each parameter if all parameters have default values which are pre-set at the outset of learning; for convenience of discussion we will assume default values in what follows. For *n* (independent) binary parameters, there would be 2^n distinct grammars for the learner to choose between, but at most *n* observations to be made; e.g., no more than 30 facts to be observed to set 30 parameters licensing over a billion grammars.[9] Chomsky implicitly assumed that each such fact would be simple, and readily detectable in the input available to the learner. Once set, a parameter value could have far-reaching and complex consequences for the structure of the target language (e.g. Hyams (1983)), so the learner’s observation would not constitute evidence in a traditional sense; rather, it functions just as a ‘trigger’ that causes the value to be adopted by the learner.

The radical simplification of what a child had to learn implied a corresponding simplification of the learning process, without need of any linguistic reasoning, generalizing, or integration of patterns across multiple sentences. The triggering of parameters was viewed as more or less automatic, effortless and instantaneous, offering an explanation for how children acquire the target language in just a few years, during which their ability to acquire other complex skills (shoe tying, multiplication) typically progresses more slowly and haltingly.

This P&P model was widely embraced by linguists and psycholinguists, at least those working within the Chomskyan tradition of generative grammar, which will be our main focus here. Nevertheless, it was a decade before any parameter-setting model was computationally implemented. This delay was certainly not due to any lack of interest. Perhaps the P&P learning model was perceived as so simple and mechanical that implementing it computationally was hardly necessary. Alternatively, with hindsight, one might surmise that attempts to embody it had discovered – sadly – that parameter setting cannot be automatic and effortless after all. Instead, as we now explain, it becomes ensnared by many of the complexities of human language that it was designed to escape.

We will review here three notable implementations of parameter setting developed during the 1990s. While they differ from each other in various ways, they all depart from Chomsky's original conception of triggering, in that they portray learning as a process of testing out whole grammars until one is found that fits the facts of the target language. Instead of the *n* isolated observations needed to trigger *n* individual parameters, these approaches require the learner to search through the vast domain of 2^n grammars. In this respect these models resemble a Gold-type enumeration-based learner, although they employ other techniques for guiding the search, borrowing heuristics from computer science that were originally developed for purposes other than modeling language acquisition. Persuasive reasons were given at the time for this turn toward domain search, as we now review.

11.5.1 Clark's genetic algorithm

A series of papers by Clark (1990; 1992) first sounded the alarm that the linguistic relationships between parameter values and their triggers are often quite opaque. Clark presented examples of parametric ambiguity: sentence types which could be accommodated by re-setting either one parameter or another. He noted that an accusative subject of an infinitival complement clause (e.g., *John believes me to be smart*) could be ascribed to a parameter permitting exceptional case marking (ECM) by the matrix verb *believes*, or to a parameter permitting structural case marking (SCM) within the infinitival complement clause. He also pointed to interactions between different parameters, illustrated by an interaction between ECM/SCM and a binding domain parameter. With a negative value for the ECM parameter, an anaphor as subject of the complement clause (e.g., *John$_i$ believes himself$_i$ to be smart*) would demand a positive setting of the parameter licensing long-distance anaphora (LDA), but with a positive value of the ECM parameter (as is correct for English) it would not imply LDA. Ambiguities and interactions such as these create a minefield in which the learner might mis-set one parameter, which could distort the implications for another parameter, creating more errors. Over all this hangs the danger that some mistakes might lead to improper adoption of a superset grammar, from which no recovery is possible without corrective data.

The conclusion that Clark drew from these important observations was that parameter setting cannot be a high-precision process but must involve a considerable amount of trial and error testing. Moreover, what must be tested is whole grammars, to avoid having to disentangle the interacting effects of individual parameters. The specific mechanism that Clark (1992) proposed drew from the computational literature on ‘genetic algorithms’, applying their domain search procedures to the case of natural language. His genetic algorithm (GA) for syntax acquisition tested multiple grammars on each input sentence, ranking them for how well they succeeded in assigning a coherent syntactic structure to the word string. Each grammar in a batch (a ‘population’) is associated with a parsing device which is applied to the sentence, and the resulting parse is evaluated by a ‘fitness metric’, which values economy of structure and which penalizes analyses that violate UG principles or the Subset Principle. After evaluating one population of grammars in this way, the GA would repeat the process with another and another, and would begin to ‘breed’ the highest-performing grammars from each. That is, their parameter values would be mixed into new combinations, creating new grammars which could then be evaluated in turn, in hope that eventually just one grammar would stand out as superior to the others, while low fitness grammars are removed from the pool. Clark also assumed a ‘mutation’ operation which changes the value of a parameter at random, as a means of introducing fresh variation into the pool.

The GA’s search for the target grammar is thus linguistically directed, with the pool of candidate grammars being progressively narrowed to those that best cover the input data. As Clark observed (1992: 95): “it should be clear that a learner cannot blindly reset parameters. He or she must, in some sense, be aware of the potential effects of setting parameters and must have some means of moving toward the target, based on experience with the input text.” However, the amount of work involved in extracting relevant information from the input text is clearly far greater than was originally anticipated by the P&P theory. The GA keeps track of large numbers of grammars, and obtains rich information about the sentence structures assigned by their associated parsers. Inevitably the question arises as to whether these operations would exceed the computational capacity of a small child. It is not easy to imagine that each time a toddler hears a sentence uttered by a parent, s/he attempts to parse it with many different grammars in the small amount of time before responding to that utterance or hearing another one. Also, and surprisingly in view of the substantial memory and computing powers ascribed to this learner, Clark reports that the GA fails to acquire the correct parameter values in some cases. This is because the grammar-breeding process is guided by whole-grammar fitness, and is only approximately related to the correctness of the specific parameter values comprising those grammars. So the GA will sometimes favor grammars with an incorrect parameter value. If the correct value of that parameter is not present in the narrowed pool of highly valued grammars, then it will be difficult to recover it; only the mutation operator could do so, and the probability that it would target the one parameter in question is low. Clark notes that this danger grows as the pool of candidate grammars is whittled down in later stages of learning, i.e., as the learner is closely approaching the target grammar. This is called the 'endgame problem' and is well-known in studies of genetic algorithms for other types of learning besides language. Making a virtue of this, Clark and Roberts (1993) observe that the susceptibility of GAs to mis-convergence can be an advantage in accounting for language change, as has been argued for other models as well (e.g. Lightfoot (1991)).

11.5.2 Gibson & Wexler's Triggering Learning Algorithm

In a much-cited article, Gibson & Wexler (1994) presented a parameter setting model which they called the Triggering Learning Algorithm (TLA). Like Clark's GA, the TLA did not reflect Chomsky's picture of 'automatic' parameter setting in which an encounter with a trigger datum in the input would simply flip the appropriate parameter switch to the correct value. Despite its name, the TLA had no knowledge of what the triggers are for the various parameters, and hence no means of detecting them in the input. Most specifically, it had no means of identifying which parameters could be re-set in order to license a particular novel input sentence. The TLA was input-guided only in the minimal sense that, once a grammar hypothesis had been selected by the learner for testing on an input sentence, that sentence could support or disconfirm the hypothesis.

The TLA proceeded by trial and error, making random guesses about what parameter values to try out, constrained by three general procedural constraints: change the current grammar only if it has failed to license the current input sentence; change to a new grammar only if it licenses the current input sentence; change the value of at most one parameter in response to one input sentence. The first of these ('error driven learning') was intended to provide stability in the sequence of grammar hypotheses; the second ('Greediness') was intended to yield improvement over the past grammar hypothesis; the third (the 'Single Value Constraint') was intended to conserve as much past learning as possible while updating the grammar in response to an as yet unlicensed input.

A major advantage of the TLA over the GA, with obvious relevance for psychological feasibility, is that the TLA places minimal demands on memory and processing capacity. Like the GA, the TLA uses the sentence parsing mechanism to try out a grammar on an input sentence. But where the GA tests multiple grammars on the same sentence, the TLA tests only one new grammar at a time, and then only if parsing with the current grammar has failed. Also, where the GA extracts rich information from each parse (e.g., the number of grammatical principles violated, the lengths of syntactic chains), the TLA receives only a simple notification of success/failure of the parse. Thus, the TLA represents an opposite extreme from the GA, employing just about the simplest imaginable mechanism. It walks around the domain of possible languages, one small step at a time, seeking to arrive at the target just by turning away from unsuccessful grammars and toward more successful ones.

A likely consequence of reduced computational power is reduced success, and indeed G&W showed that the TLA would fail to converge on the target grammar in a number of cases (see below). The problem is caused by what are known as *local maxima* (sometimes *local* *minima*), in which the learner's current hypothesis is wrong but the only other accessible hypotheses (limited by the Single Value Constraint) are no better. When the current hypothesis is *locally* the best one available, this learning mechanism cannot escape to explore other areas of the grammar domain which might afford even better hypotheses. This is not unrelated to the 'endgame' problem of Clark's GA. It is a well-known property of the class of 'hill-climbing' algorithms, which search a hypothesis space by making local changes in the direction of improvement. Such searches are too short-sighted to be able to detect an overall optimal hypothesis beyond the neighborhood currently being scoured.

It is unclear just how great a price the TLA pays for its extreme economy of resources. But Gibson & Wexler (1994; henceforth G&W) reported that the TLA failed to converge on the correct parameter values for 3 out of 8 languages in a tiny artificial language domain defined by 3 binary parameters (initial/final subject, initial/final head in VP, +/-Verb Second). Niyogi & Berwick (1996) subsequently provided a more detailed probabilistic analysis demonstrating additional TLA failures. By contrast, Turkel (1996) showed that a GA did not succumb to any local maxima errors on this same domain. It is possible that the properties of the G&W 3-parameter domain are not characteristic of the far more extensive domain of natural languages. It can be hard to predict in general whether performance would be improved on a larger domain, due to richer input information, or would decline due to more parameter interactions. However, increased domain size did not assist the TLA when it was tested on a 12-parameter domain (Kohl (1999)). It was found that only 57% of the 4,096 languages in that domain could be reliably attained without the TLA getting trapped in a local maximum, even given the most facilitative starting-state default parameter values.[10]

G&W explored various ways of avoiding these failures (e.g., giving up the SVC and/or Greediness constraints). Noting that the problematic target grammars in the 8-language domain were all –V2 grammars, G&W contemplated designating –V2 as a default value, hence needing no triggers. The problem with this solution, for a non-deterministic (trial-and-error) system such as the TLA, is then to make sure that the default value is not prematurely given up by mistake, which could spark further errors. (See Bertolo (1995a,b) for discussion of possible remedies.) This model is not being actively pursued these days; however, a model akin to it but more resilient to local maxima has since been introduced by Yang.

11.5.3 Yang's Variational Learning Model

The new domain-search model developed by Yang (1999), Yang (2002) have some of the strengths of both GA and TLA models. Yang's Variational Learning model (VL) employs the more modest processing resources of the TLA, testing only one grammar on each input sentence, but like the GA it has a memory that registers a measure of past successes. In the case of the VL (unlike the GA), parameter values are assessed individually. As usual, parameters are assumed to be binary, with values 0 and 1. In the memory, a scale is associated with each parameter, and a pointer on the scale indicates the current weight of evidence for one value or the other. Whenever the 0 value of some parameter P is in a grammar that successfully parses an input sentence, P's pointer is moved a small amount in the direction of the 0 value (i.e., the 0 value is 'rewarded'); likewise, the 1 value of P is rewarded whenever it is part of a grammar that just succeeded in parsing an input. When the 0 value or the 1 value of P is part of a grammar that just failed to parse an input, that value is 'punished', by moving the pointer a small way away from that value, towards the opposite value. If the pointer comes to within a very small margin from the 0-end or the 1-end of the scale, the parameter is deemed to have been set permanently to that value.

When the VL encounters a new sentence, it selects a grammar to try parsing it with. The selection is probabilistic, based on the weighted scale-values of the various parameters in the domain. Grammars composed of previously successful parameter values are more likely to be selected for testing on a new input. But even low-valued grammars have some chance of being selected occasionally. Thus the VL benefits from its past experience, as it accumulates evidence of which parameter values seem to suit the target language; but because it also samples a wide variety of other (less highly valued) grammars, it does not risk getting locked into an enticing but incorrect corner of the grammar space, as the TLA did. The occasional testing of low-valued grammars thus serves a similar purpose to the GA mutation operator in providing an escape route from local maxima.

Yang has reported positive results in simulation experiments with the VL. In one study he notes that "the learner converges in a reasonable amount of time. About 600,000 sentences were needed for converging on ten interacting parameters" (Legate and Yang (2002: 51)). A formal proof for that version of the VL shows that it is guaranteed to converge in any parametric domain except those containing subset-superset relationships (Straus (2008)). However, Straus also proves that in order to converge in some possible domains, the VL could consume a number of sentences exponential in the number of parameters. For any reasonable number of natural language parameters, learning would clearly be impracticable in this worst case. Yang (2012) present a 'reward only' variant of the model, which is shown to perform more efficiently than the original reward-and-punish version, especially in a language domain "favorable to the learner", i.e., with abundant unambiguous triggering data.

The VL has psycholinguistic merits. Gradual learning is characteristic of the VL's performance in the simulation studies, and Yang points to the child language acquisition literature as showing that a child's transition to the target value of a parameter is gradual, with no sharp changes such as a classic 'switch-setting' model would imply. Also, the order of the VL's acquisition of different parameters is correlated with the frequency of evidence for them in children's linguistic environment. Yang (2011) notes that this too mirrors children's linguistic development: for seven familiar parameters (including, e.g., Wh fronting and verb raising) the time course of child acquisition is predicted by the amount of evidence available in child-directed speech. This commitment to matching the performance of an implemented computer model to the facts of child language development establishes an important goal for all future modeling endeavors.

But the VL has some quirks as a psychological model. Like the TLA it selects a grammar hypothesis to test without first consulting the properties of the current input sentence, so it may try out parameter values that bear no relation to the needs of that sentence, e.g., testing the positive value of preposition stranding on a sentence with no preposition. Moreover, that totally irrelevant preposition-stranding parameter value will actually be rewarded if the grammar it is included in succeeds in parsing that preposition-less sentence. Also, the VL predicts significant failures of sentence *comprehension* by learners. It does not always parse input with its currently most highly valued grammar, because of its need to sample lower-valued grammars. A low-valued grammar would sometimes, perhaps often, fail to parse a target-language sentence. So if children behaved in the manner of the VL, there would be occasions on which a child would be unable to parse a sentence, and hence would fail to understand it – even if were a sentence she had understood just moments before when she parsed it with her current 'best' grammar! Without hard data against which to test this aspect of the model it cannot be regarded as a fatal defect, but it does seem implausible that all normal children deliberately process language with grammars they believe to be incorrect, sacrificing comprehension in order to achieve safer learning.

11.5.4 Domain Search: Taking Stock

Three major approaches to modeling syntax acquisition, outlined above, follow Chomsky's lead in assuming that natural language grammars consist of innate principles and parameters, with only the parameter values needing to be established by the learner. But they all depart from Chomsky's original conception of acquisition as the triggering of individual parameters in a process that is psychologically plausible (resource-light), but is accurate and fast due to being strongly input-guided: on encountering a sentence of the target language, not yet licensed, the learning device would have direct knowledge of which parameters could or must be reset to accommodate it. The GA is the closest to being input-guided but is not resource-light. The TLA and VL are resource-light but not input-guided. For reasonably sized natural language domains, none of them is fast, though the VL is accurate in the weak sense that it is guaranteed to converge on the target grammar eventually. Nevertheless, however much these models differ from classical triggering and from each other, they do all come to grips with the twin problems of parametric ambiguity and parametric interaction, as emphasized by Clark. They can survive parametric ambiguity because they are non-deterministic, so any temporary mis-settings do no permanent harm (as long as they do not lead to Subset Principle violations or local maxima). These models can also survive complex parametric interactions because they do not even attempt to establish the values of single parameters in isolation from other parameters in a grammar. In one way or another, they assess whole grammars rather than choosing between the two values of each of n parameters – thereby undermining the economical and widely heralded "Twenty Questions" character of Chomsky's P&P proposal. These models have the advantage of drawing on known computational techniques, showing how general purpose learning mechanisms can be adapted for language learning. As long as the UG principles and parameters are supplied as the knowledge base, there is no need to posit innate learning procedures specialized for human language. This is in keeping with the current emphasis on minimizing the complexity of the innate component of the language faculty (Hauser, Chomsky & Fitch 2002), by shifting the burden to general cognitive mechanisms. In particular, unlike the original switch-setting model, these learning models do not need to be innately equipped with *knowledge of the triggers* for the parameter values. This will be relevant to discussion below.

On the other hand, as noted above, the scale of domain search over the full domain of natural languages is immense, and non-deterministic techniques multiply that problem by repeating some steps many times over. Linguists have offered different estimates of how large the domain of natural languages is. Early estimates that about 30 syntactic parameters (already yielding over a billion grammars if parameters are mutually independent) would suffice to codify all cross-language syntactic differences, have given way in recent years to estimates many times larger. There may be several hundreds or thousands of micro-parameters (Kayne (2005a)), which explode exponentially into trillions of trillions of candidate grammars. Promising proposals have been developed for systematizing and constraining the class of possible parameters (Gianollo, Guardiano & Longobardi (2008); Roberts (2007, 2008), Roberts & Holmberg (2007)) and for ordering them hierarchically so that some can be disregarded until others have been set Villavicencio (2000); Baker (2008b); Biberauer, Holmberg, Roberts & Sheehan (2010)). As a simple example, a parameter for partial wh-movement is irrelevant until a higher level parameter licensing wh-movement of any kind has been set to its positive value. If such proposals succeed in radically limiting the set of parameters that must be consulted at any stage in the acquisition process, then there may be a future for search-based learning models. But it is unclear at present whether the problem of scale can be tamed sufficiently to permit a search model to converge without exceeding either the cognitive resources of child learners or the time-frame of their progress.

11.6 Implementing parameter setting: input guidance

Despite the preponderance of trial-and-error domain-search learning models during the 1990s, theoretical linguists were not greatly shaken in their allegiance to the original approach to parameter setting as envisaged in the 1980s, such that an input sentence not yet licensed by the learner's grammar could *reveal* to the learning mechanism, without resort to extensive trial-and-error, which parameter settings could license it. Direct guidance by the input as to which parameter(s) it would be profitable to reset still seemed to hold the greatest promise of rapid and accurate learning. Learning theorists' doubts about its feasibility did not deter syntacticians from continuing to propose unambiguous triggers for various parameters over the years (notably Hyams (1983, 1992) for the Null Subject parameter; Baker (1996) for the polysynthesis parameter; Pollock (1997) for the verb raising parameter). This optimistic perspective in theoretical syntax circles might have been expected to spur the development in computational modeling circles, of input guided parameter setting systems in which Chomsky's switch setting metaphor would be concretely implemented. In fact, it was phonology that led the way.

Dresher and Kaye (1990) pioneered an input-guided learning system for a collection of phonological parameters responsible for stress assignment, such as bounded vs. unbounded foot size; extrametrical syllables permitted at right or left (Dresher & Kaye (1990); Dresher (1999);

henceforth DK). DK were well aware of the challenges involved. They noted that this approach requires the learning model to have knowledge of the triggers (which they referred to as "cues") for all the parameters, and that these cues must be fully reliable, not infected by the ambiguity and interaction problems discussed above. This is made more difficult by the fact that the linguistically relevant triggers are typically abstract properties not directly observable in the raw input a learner receives. In terminology introduced by Chomsky (1986b), the perceptible input is E-language, but the learner needs to extract abstract I-language facts from it.[11]

DK were able to solve problems raised by interactions among the metrical parameters they studied by imposing an ordering on the parameters, such that a reliable cue for one parameter could be discerned only once some other parameter(s) had been set correctly. The importance of an orderly sequence of learning events is reflected in the title of Dresher's 1999 paper "Charting the learning path". The model also makes extensive use of defaults, both within parameters and between parameters, to establish priorities for the analysis of ambiguous inputs (e.g. in the absence of an unambiguous cue, assume the value Quantity-insensitive value of the Quantity Sensitivity parameter). Some thorny problems remained, such as the unreliability of parameter values established on the basis of the *default* value of some other parameter, in case that default value was later overthrown by new input. One solution, adopted by DK in their 1990 implementation YOUPIE, was to assume non-incremental ('batch') learning, i.e., collecting up all relevant data before setting any parameters (contra psychological plausibility). An alternative, embraced by Dresher (1999), was to set back to their default values all parameters that had been set on the basis of the default value now overthrown. Another matter demanding reflection was how a learner could be constrained to adhere strictly to the learning path, without which there would be no way to avoid entanglement in troublesome parametric interactions. Assuming that infants could not discover the correct path unaided, it was presumed that it must be dictated by

UG, along with the universal principles and parameter values and their cues. This was not incompatible with early views of the richness and specificity of innate linguistic knowledge, but it clashes with current attempts to minimize the extent to which biological specialization for language must be assumed (Hauser, Chomsky & Fitch 2002).

For syntax, one might set about developing an input-guided model of parameter setting in a similar fashion, by inspecting the parameters and triggers proposed by syntacticians on linguistic grounds, and attempting to tidy away any damaging ambiguities or interactions between them. Following DK this might be achieved by refining the triggers, by imposing default values, and/or by ordering the parameters. Any such devices discovered to be necessary to ensure convergence on the learner's target language might then be posited as part of the learner's innate endowment or, where possible, be attributed to what are now called 'third factor' influences, not specific to language, such as least effort tendencies or principles of efficient computation (Chomsky (2005)).

We have made a start on this project, working with a modest collection of 13 familiar GB-style syntactic parameters in the context of a "structural triggers" learning model (Fodor (1998a); Sakas & Fodor (2012)).[12] In this model the parameter values are taken to be UG-specified I-language entities in the form of 'treelets'. A treelet is a sub-structure (a collection of syntactic nodes, typically underspecified in some respects) of sentential trees. A simple example would be a PP node immediately dominating a preposition and a nominal trace, signifying that a language can strand prepositions (as in English *Who did you play with today?*, as opposed to pied-piping in standard French *Avec qui avez-vous joué aujourd'hui?).*

The motivation for parametric treelets is that they can be used by learners to rescue a parse of a novel input sentence which has failed for lack of a needed parameter value; and in doing so they can *reveal* what new parameter value is needed. It is assumed on this approach that a child's primary aim is to understand what others are saying. So the child tries to parse every sentence, using whatever her currently 'best' grammar is. If that succeeds, the child is doing just what an adult would do. But if it fails – i.e., if at some point the parse tree can't be completed on the basis of the current grammar, the learning mechanism then consults the store of parametric treelets that UG makes available, seeking one that can bridge the gap in the parse tree. If some treelet proves itself useful, it is adopted into the learner's grammar for future use in sentence processing (or, if learning is gradual, the activation level of that treelet is slightly increased, making it incrementally more accessible for future sentence processing).

For example, imagine a learner of English who is familiar with wh-question formation but has not yet acquired preposition stranding (Sugisaki and Snyder (2006)), and who hears the sentence *Who did you play with today?*. The child would process this just as an adult does, up to the incoming word *today*. At that point the child's current grammar offers no means of continuing; it contains no treelet that will fit between *with* and *today*, and so *today* cannot be attached into the tree. The parser/learner will then search for a treelet in the wider pool of candidates made available by UG, to identify one which will fill that gap in the parse tree. Since *with* heads a prepositional phrase, relevant UG treelets to be considered are those dominated by a PP node; among them, a PP with a null object will allow the parser to move forward to the word *today*. In this way the specific properties of input sentences provide a word-by-word guide to the adoption of relevant parametric treelets, in a narrowly channeled search process (find a treelet containing a preposition not followed in the word string by a potential object).

A child's sentence processing thus differs from an adult's just in the need to draw on more treelets from UG because the child's current grammar is as yet incomplete. In other respects the child is relating to language just as adults do. No additional language acquisition device (LAD) needs to be posited. The parsing mechanism exists in any case, and is widely believed to be innate, accessible to infants as soon as they are able to recognize word combinations to apply it to. And the tricky task of inferring I-language structure from E-language word strings is exactly what the human sentence parsing mechanism is designed to do, and does every day in both children and adults; the transition from learner to fully competent adult speaker is thus seamless.

In comparison with the domain search models discussed above, this structural triggers learning-by-parsing approach has some efficiency advantages. It *finds* a grammar that can parse a novel input, unlike trial-and-error learners which first pick a grammar on some other basis and then find out whether or not it succeeds. So it avoids wasting effort trying out arbitrarily chosen parameter values, or testing grammars that fail, and it avoids the loss of comprehension that such failures would cause. Learning by parsing thus predicts faster convergence (confirmed in simulation studies, e.g., Fodor and Sakas (2004)) and fewer errors of 'commission' (Snyder (2011)) en route to the target. It also avoids the distracting free rides in which a parameter value is strengthened regardless of whether or not it contributed to a successful parse. Like the domain search models, it is compatible with the eliminative ambitions of recent linguistic theorizing, since it posits no more mechanism than is inherent in the ability to produce and understand language at all. It is also a potential source of third-factor influences since the parser exhibits some economy tendencies in familiar parsing strategies such as Minimal Attachment and the Minimal Chain Principle, as well as being open to frequency influences.

The major cost of this approach is that the parametric treelets must be innately specified. How burdensome that is, and how plausible it is from an evolutionary point of view, remains to be determined. Note, however, that the treelets do double duty as (a) definitions of the parameter values themselves and (b) I-triggers for detecting those parameter values in the input.[13] Every parametric theory must do at least the first, i.e, say what parameters it assumes. E-triggers do not need to be innately specified on this approach since they follow from the interplay of the I-triggers with general UG structural principles in the course of parsing; a parameter value at the I-level is acquired just in case it solves an on-line problem that the learner/parser has encountered at the E-level. Another uncertainty at present is how the local treelets that would be efficient for parsing relate to the derivational representations of recent linguistic theory, such as the bottom-up derivations resulting from Merge (but see Chesi (2014)).

11.7 Learnability theory and linguistic theories

The trajectory of this survey reflects the course of thinking over the past 50 years, primarily within the tradition of transformational generative grammar, about the nature of human language and how its breadth and intricacies could be acquired by children so accurately and rapidly. Modern linguistic and learnability theory dawned with the riveting discovery that natural languages have a formal structure that can be studied with mathematical rigor. The emphasis today has shifted to a conception of the language faculty as a biological organ, functioning in concert with other perceptual and conceptual systems. The learning-by-parsing approach to parameter setting discussed above fits well with this latter stance, so it is worth observing that learning by parsing is compatible with a wide swath of current linguistic theories, even if they diverge greatly with respect to how they characterize the ways in which natural language grammars can differ. The structural elements that vary from one language to another, whether they are called parameters or not, are what the learner must choose between.

Theoretical linguistic frameworks can thus be broadly characterized from a learnability perspective by examining the tools they offer to a parser/learner. One obvious difference between linguistic theories is the size of their 'treelets' – the structural building blocks of sentential tree structures which are assumed to differentiate natural languages. The Minimalist Program (MP; Chomsky (1993, 1995b, 2001a) and since) represents the lean extreme on the scale of treelet size. The active elements that drive the course of a syntactic derivation are individual formal features (e.g., case features, agreement features) on functional heads. The 'probe-goal' apparatus which is the driving force of MP derivations establishes Agree relations between pairs of syntactic elements (a probe and a goal) in a tree structure, under which they can supply values for each other's unvalued features. In order for this to occur, a goal constituent must in some cases, but not all, move to the neighborhood of the probe, thereby creating cross-language word order differences. Whether or not movement is triggered is determined by whether or not the probe carries a strong feature (or EPP feature or edge feature). Significant details are omitted here, but the import of this aspect of the Minimalist Program for syntax acquisition is discernable. In place of the free-standing parameters of the Government-Binding theory tradition, cross-language syntactic variation in MP is controlled by the featural composition of certain specific lexical items, some of which may be phonologically null. What is needed for acquisition is then a parser which can apply such a grammar on-line, and which can extend the learner's inventory of formal features over time, as needed to build structure for the sentence patterns encountered in the input.

By contrast, the structural units of cross-language variation in other theoretical frameworks can be quite large. In Tree Adjoining Grammars (TAGs), for example, the basic ingredients of syntactic structures are specified by the grammar as 'elementary trees', which may span a complete clause. UG defines the set of possible elementary trees, and the adjunction and substitution operations by which they can be combined (Frank (2004)). Construction Grammar, in its several variants[14], also admits syntactic building blocks with clausal scope. In the Cognitive

Construction Grammar of Goldberg (2005), languages differ only with respect to which constructions they admit, where a construction is an abstractly characterized pattern that pairs aspects of form with a distinctive semantic or discourse role. Familiar examples are the passive construction, the subject-auxiliary inversion construction, the ditransitive construction, which are integrated in a sentence such as *Will Mary be given a book?*.

Other aspects of linguistic theories which may have consequences for learnability include how numerous their (counterparts of) structural 'treelets' are; how diverse they are; how surface-recognizable they are; and what, if any, general principles constrain them. In the present stage of research it is unclear which will prove to be the most important grammar characteristics for learnability, but to hazard a guess, they may well be: Does the grammar format afford an effective psychological parser? (This is uncertain but under study at present for MP.) How deep or shallow is the relation between E-language and I-language structures? (Construction Grammar might anticipate an advantage for shallower I-structures.) Do subset and superset grammars stand in some systematic formal relation which could be exploited by learners to avoid SP violations? (Sadly, no linguistic theory has this property as far as is known now.)

Not all current linguistic theories have been the target of simulation studies or formal theorems to test their potential for accurate and efficient acquisition in the way of the parametric models discussed in this article. But there is clearly worthwhile learnability research work to be done on all theoretical approaches. Psycholinguists are ever hopeful of finding a performance measure which differentiates between linguistic theories and can proclaim one more explanatory than others. It is possible, and to be hoped, that learnability as a litmus of the psychological reality of formal linguistic theories will one day play a more significant role than it has to date.

[1] There was also a rich tradition in categorial grammar research beginning with Ajdukiewicz (1935), continuing with the work of Bar-Hillel (1953) and Lambek (1958). For recent work in this tradition see Steedman and Baldridge (2011), and references there.

[2] Note that for convenience from now on we refer to subset relations between *grammars* as a shorthand for subset relations between the *languages* (sets of sentences) that the grammars generate.

[3] Since Gold's enumeration learner was able to eliminate many grammars on the basis of a single new input sentence, this would not necessarily cause significant delay. Translated into psychological terms, however, this would amount to parallel processing of multiple grammars, on a possibly massive scale, which would disqualify it on grounds of psychological feasibility.

[4] This assessment was made before parameter theory (Chomsky (1981)) imposed a finite limit on the number of possible human grammars, with the expectation of a finite number of learning steps to acquire them. But problems of scale have nevertheless remained central to learnability research; see below.

[5] Pinker warns that the heuristic approach has its own pitfalls, however, if what is postulated is a large and unruly collection of ad hoc heuristics.

[6] "… a transformation can involve material only in the S on which it is cycling or the next S down" (W&C, p. 310).

[7] "If the immediate structure of a node in a phrase-marker is nonbase, that node is *frozen*."; "If a node A of a phrase-marker is frozen, no node dominated by A may be analyzed by a transformation." (W&C, p. 119).

[8] Parameters were at first regarded as variables in the statement of the constraints, e.g. specifying whether S or S' counted as a bounding node for subjacency (Rizzi (1982)). In later work, parameters were associated with lexical items, especially functional heads, e.g., the head-direction parameter could be recast in terms of the direction of government of a given head (V, P or the Infl(ection)/T(ense) morpheme); see in particular Travis (1984), Koopman (1984) and Chapter 14.

[9] For estimates of the actual number of parameters, see section 5.3 below.

[10] Also, Fodor & Sakas (2004) observed 88% non-convergence by the TLA on a 13-parameter domain of 3,072 languages unrelated to Kohl's domain.

[11] The terms I-language and E-language were coined by Chomsky (1986b: 19-22). I-language is "internalized language" which, following Jespersen, is "some element in the mind of the person who knows the language." This

contrasts with E-language, which is "externalized language", i.e., following Bloomfield, "the totality of utterances that can be made in a speech community." See further discussion below.

[12] Variants of this model can be found in Fodor (1998b), Fodor & Sakas (2004). Sakas (2000) and Sakas and Fodor (2001) present a more computational exposition of some of the issues.

[13] Lightfoot (1991) also emphasizes the I-language status of syntactic triggers.

[14] Goldberg (2005, Ch.10) outlines similarities and differences among several theories which embrace constructions as a (or the) major descriptive device for characterizing natural languages.