Sakas, W.G., Berwick, R.C and Yang, C.D. (to appear) *Linguistic Analysis.*
Pre-publication draft: Please do not quote without permission

Parameter setting is feasible

William Gregory Sakas
Hunter College and The Graduate Center
City University of New York

Charles Yang
University of Pennsylvania

Robert Berwick
Massachusetts Institute of Technology

1. Introduction

In the first few years of life, a child's language undergoes tremendous changes. While it is a primary task for the student of language acquisition to document these changes, a mere description of child language, however accurate or insightful, cannot be regarded as adequate. A complete theory of language acquisition must also include a detailed account of *how* these changes take place, that is, the mechanism of language learning. The study of this mechanism must take into account contributions from both the internal knowledge of language, innate or otherwise, and the external linguistic experience, which determines the outcome of learning. The goal of language acquisition research is to establish the right combination of the theory of grammar and the theory of learning such that the development of language can be successfully accounted for.

Viewed in this light, the Principles-and-Parameters (P&P) framework (Chomsky 1981) represented a radical shift of paradigm for both the theory of grammar and the theory of learning. First, as more and more languages were subjected to generative studies, a number of universal principles emerged, ones which are not restricted to specific constructions or particular languages. Second, a great variety of sentence structures can be efficiently described by a small number of parameters; different grammars are instantiations of different operational choices in a universal engine of sentence building, much like configuring computer software. The principles, which are putatively innate and universal, are not learned, and can be expected to be operative in (early) child language; this opens up a wealth of topics for empirical research. On the other hand, the parameter values, which vary cross-linguistically, must be learned on the basis of specific linguistic evidence, which also can be quantified and evaluated empirically (Yang 2002). Thus, the commonalities and differences in children's acquisition of specific languages receive a principled and unified interpretation. In the original conception, parameters are like data compression devices, designed to be anchor points for dividing up the linguistic space: the complex interactions among them would provide coverage for a vast array of linguistic data—more "facts" captured than the number of parameters, so to speak—such that the determination of the parameter values would amount to a simplification of the learning task. This conceptual notion of parameters goes well with the perspective of machine learning and statistical inference, where plausible learnability can only be

achieved by constraining the hypothesis space within some finite dimensions (Valiant 1984, Vapnik 1995, see Niyogi 2006 for review). Moreover, if the number of parameters is finite, then there is only a finite--albeit large, perhaps--number of grammars that forms the child's learning space; this sidesteps the well-known problem of inductive indeterminacy in an infinite hypothesis space associated with phrase structure rules (Gold 1967, Chomsky 1981).

That is not to say that the problem of parameter setting, and that of language acquisition, is resolved. Many commentators have remarked on the sheer number of grammars that are made possible by the combination of parameter values, which will surely defeat a brute-force exhaustive search (Berwick & Niyogi 1996). But the number of grammars itself is not the core issue; what determines the complexity of learning is the *structure* of the parameter space (Sakas, 2000). Even a space of infinitely many hypotheses can be easily learned if it is "smooth" so as to allow for efficient search.[1] For instance, there are infinitely many linear functions in a 2-D space but a handful of points drawn from it suffices to establish the slope and intercept of the target linear function, which is after all the process of linear regression. In this sense, the space of linear functions is structurally simple such that the target hypothesis can be effectively established.

In the first two decades of the P&P framework, there was considerable effort devoted to the study of parameter setting and its implications on language acquisition (Berwick 1985, Roeper 1987, Kapur, 1993, Gibson & Wexler 1994, Fodor 1998a, 1998b, Dresher 1999, Sakas & Fodor 2001, Yang 2002, Fodor & Sakas, 2004). Most of these learning models, however, operate on toy domains although even a simple three-parameter system studied by Gibson & Wexler (1994) proved much harder to learn than one might have expected (Berwick & Niyogi 1996). At the same time, the past fifteen years or so have seen the parameter falling on hard times as alternative conceptions of how to encode cross-linguistic differences have begun to (re)emerge (Newmeyer 2004, Culicover & Jackendoff 2005).[2] It is certainly logically possible to recast the fact of language variation without appealing to syntactic parameters; we can point to variation in the lexicon, variation in the functional projections, features, feature strengths, feature bundles, etc. to "externalize" the parametric system to interface conditions, presumably out of the syntactic system proper. But it is also important to realize that such a move does not fundamentally change the nature of acquisition problem: the learner still has to locate her target grammar in the space of finite choices with a reasonable amount of data within a reasonable amount of time. And to the extent that syntactic acquisition can be viewed as

[1] By "smooth" here we mean any structure of the hypothesis space that allows for efficient search. Smoothness is sometimes used as a more technical term to mean a correlation between the similarity of grammars and the languages they generate (Sakas, 2000; Berwick & Niyogi, 2996). A domain in which there is such a strong correlation is one type of smooth domain in the current sense of the term.

[2] Although some of these challenges are mistakenly directed toward the parameters, on our view. For instance, Newmeyer (2004) is concerned by the fact that the parameter values of a language may have exceptions for certain lexical items or constructions. But that observation merely calls for a theory of learning that can identify and tolerate exceptions (e.g., Yang 2005, 2016): a rule-based approach will also have exceptions and abandoning parameters does not seem to change the nature of the problem.

a search among a constrained set of grammatical possibilities, as reviewed above, the Minimalist—and indeed, non-Minimalist—alternatives to parameters ought to provide similar empirical coverage.

We still believe that the parameters are the best solution for the problem of language acquisition but the value and benefit of this approach must be demonstrated. In this paper, we aim to move beyond toy grammars and provide a large-scale study of parameter setting in a linguistically complex domain constructed by the research group at The Graduate Center and Hunter College of the City University of New York CUNY-CoLAG (Sakas and Fodor, 2011, 2012). Section 2 presents several prominent parameter setting models, how how they deal with the problem of parametric ambiguity: multiple collections of parameter values can license the same input sentence. These models include the Triggering Learning Algorithm of Gibson & Wexler (1994), the Structural Triggers Learner of Fodor (1998a) and Sakas and Fodor (2004), and the probabilistic variational learning model of Yang (2002). Section 3 provides a concise description of the parameter domain which has implemented 13 syntactic parameters that are largely used for encoding cross-linguistic word order variation. In Section 4 we show that with the exception of the trigger model of Gibson & Wexler, these learners are able to navigate a complex syntactic parameter space and arrive at the target grammar consuming a reasonable number of input sentences. These findings suggest that parameters, and learners that set parameters to their correct values, together form a compelling explanation of how human children learn language.

2. The Learning Models and the Problem of Ambiguity

The *Triggering Learning Algorithm* or *TLA* (Gibson & Wexler, 1994) is appealingly simple. The learner tests the current input sentence against its current grammar hypothesis G_h if G_h licenses the input the learner does nothing. If G_h doesn't licenses the input the TLA randomly picks one parameter to toggle its value (i.e., from 0 to 1 or from 1 to 0) which creates a new hypothesis G_w that differs from G_h by only one parameter value. If G_w licenses the input then G_w becomes the current hypothesis, otherwise G_h remains the current hypothesis. The learner then waits for a new input.

The TLA embodies some (apparently) psychologically desirable features. Changing only one parameter (dubbed the *Single Value Constraint* by G&W), combined with changing the hypothesized grammar only in the event that the new grammar can parse the input (*Greediness*) promotes conservatism jumping between radically different grammars is prohibited. The TLA also requires very little in the way of computational resources. It requires no memory for past sentences or grammars, and at most two grammars are applied to a single input sentence.

The *Structural Triggers Learner* or *STL* (Fodor 1998a, 1998b, Sakas and Fodor 2001, Fodor and Sakas, 2004, Sakas and Fodor, 2012) is actually a family of learners that makes use of structural information gleaned from input by the parser. In the STL model, parameters are not 1's or 0's but rather small ingredients of tree structure called *treelets*.

A treelet contains what is minimally necessary to linguistically define a parameter value. For example, a treelet for the 1 (final) value for the Subject Position parameter might be: S is the right sister of IBAR, for the 0 (initial) value: S is the left sister of IBAR. UG provides a treelet for every natural language parameter value. There is no meaningful distinction between parameter values and treelets: a treelet *is* a parameter value. The *supergrammar* (Fodor, 1998a) is the pool of all UG-supplied treelets together with universal principles. A single grammar (or grammar hypothesis) is a proper subset of the entire range of UG-supplied treelets (parameter values)—in the models being considered here a grammar consists of one treelet per parameter.

The parser applies the supergrammar to every sentence encountered. Learning consists of the strategies used in choosing which parametric treelets to employ during a supergrammar parse, and which to retain as part of the learner's grammar hypothesis—in other words which parameter values to change (if any) during the course of learning. The STL variant employed in this work is stochastic. Given a choice between a "0 treelet" and a "1 treelet" for a parameter—either of which would result in a complete parse tree—the learner picks one at random without bias, this is called the *Any Parse STL* in Fodor and Sakas (2004).[3]

Like the STL, the *Variational Learner* or *VL* (Yang, 2002) is actually a family of learners that combines domain general, probabilistic learning with UG parameters. All VL variants maintain a vector of real valued weights (0 to 1), each weight is associated with a parameter. The weights are used to guide grammar hypothesis selection during the course of learning. If a weight for parameter *Pi* is greater than .5, the next grammar hypothesis is more likely to have *Pi* set to 1, if the weight is less than .5 the next grammar hypothesis is more likely to have *Pi* set to 0. After each input sentence is consumed by the learner, the weights are updated based on a 'can parse'/'can't parse result after the learner applies the current grammar hypothesis to the input sentence. The weights effectively serve as a form of memory for which parameter values have worked best on past input sentences.

As Yang describes it (2002), the learner is effectively putting the grammars in competition with each other trying to best match the observed linguistic data. The family of VLs differ by the method used to update weights after each input sentence. The VL variant that was used in the simulations reported here is the "reward only" learner (Yang, 2012). If the current hypothesis cannot license the current input sentence, then the learner does nothing. If the current grammar hypothesis *can* parse the current input sentence, then nudge the weights up for all parameters with a 1 value, and nudge the weights down for all the 0 values.[4] Note that under this scheme, it is possible that a wrong parameter value gets rewarded as a free-rider if the overall grammar, probabilistically chosen based on the parameter weights, successfully analyzes an input sentence. The expectation is that in the long run, the parameters will gradually drift toward the target values, especially when the parameter space is structured favorably in a sense to be made precise.

---

[3] There are several other stochastic heuristics employed in this study. The family of stochastic STL models is referred to as the *Guessing-STLs*.

[4] The amount that the weights are changed follow the L_{R-P} scheme (Bush & Mosteller, 1955); see Yang (2002) for details.

It is interesting to observe how these learning models handle the problem of ambiguity. The ambiguity problem is caused by the complex interactions of the parameters: such that there are sentences that licensed by extensions of multiple grammars (i.e., combinations of parameter values). When such sentences appear the input, the learner will have multiple ways of updating their grammar hypothesis, and the challenge is to identify the correct grammar, or the path to the correct grammar, without being led astray. We now consider how the three learning models address the problem of ambiguity.

The TLA essentially ignores the ambiguity problem. If an ambiguous sentence appears, and there are multiple grammars that are one parameter value away from the current one that can analyze it, the TLA might select one randomly. Note that there is no guarantee that any of the compatible grammars will be chosen, because the TLA tries out only one new grammar upon failure with its current grammar hypothesis. Decisions are therefore strictly local: even if there is a path of grammars that eventually leads to the target, there is no guarantee that such a path will be taken. Indeed, the TLA learner often heads down the wrong path, resulting in severe convergence problems (Berwick and Niyogi 1996).

The STL and the variational learning (VL) model, by contrast, hope to sidestep the ambiguity problem if parameters have *unambiguous triggers* (Fodor, 1998, Sakas & Fodor, 2012), or *signatures* (Yang 2002).[5] Signatures or unambiguous triggers for a parameter are sentences that are analyzable only if that parameter takes on the correct value of the target language. Signatures, or unambiguous triggers we take to mean an abstract characterization of a linguistic phenomenon visible in the sentences (e.g., a non-adjacent preposition and its complement).

Strictly speaking, the VL model does not require the existence of signatures. It provably converges onto the target grammar but may takes an intractable amount of time in the worst case (Straus 2008). However, if parameters have signatures, then efficient learning becomes achievable. In particular, if parameters have signatures, then over time the probabilities of the parameters will gradually drift toward the target value. Furthermore, even if not all parameters have signatures, learning can still succeed if the parameters are *conditioned* (Dresher & Kaye, 1990; Sakas & Fodor, 2012). For instance, consider the setting of the V2 and OV/VO parameters for a language like German. Unless the OV/VO parameter is set, there is no single sentence that can set V2 to the positive value. However, once the OV/VO parameter has been set to the head-initial setting of VO, presumably based on patterns such as "participle object", and the subject has been determined to reside in the Specifier position of TP (either as a universal property or has been determined as such by the language-particular data), then patterns such as OVS becomes unambiguous evidence for the V2 parameter. The net effect is that the probability of the V2 parameter may be wandering around rather aimlessly, but it will begin a much more deterministic march toward the target value as the probability of the OV/VO is closer and closer to the target. The effectiveness of the variational learning

[5] Gibson and Wexler's (1994) *global trigger* is identical to an unambiguous trigger or a signature, see Sakas & Fodor (2012) for extensive discussion of triggers and workable definitions.

model, then, is critically dependent on the smoothness of the parameter domain: Do all/most parameters have signatures or conditioned signatures?

Deterministic STL variants ensure that the ambiguity problem does not arise by using the parser to identify unambiguous triggers. The basic idea is that any serial deterministic parser will be equipped to recognize a choice point during a parse. As the parse tree for the input sentence is constructed, if the current grammar fails, the deterministic STL adopts only those parameter value treelets that are not involved in any choice points (Fodor, 1998a). The most sophisticated deterministic STL at this writing makes use of conditioned parameters, both single-parameter and between-parameter defaults and is described in Sakas & Fodor (2012). The Any Parse STL used in this study is nondeterministic. If the current grammar fails, and a choice point is encountered, adopt a parameter value treelet randomly as long as it leads to a complete parse tree. The Any Parse STL was chosen because in some sense it is the weakest and least informed STL variant on how to deal with ambiguity. That said, it performs with surprising efficiency. This is due to the fact that all STL models perform *decoding*: STL learners know which parameters need to be reset in order for a parse to go through. This means that for parameters that have unambiguous triggers the learner will never have cause to change that parameter which causes an exponential reduction in the grammar space that needs to be searched. So in contrast to the VL model, the STL model can focus its efforts on parameters lacking triggers. For further discussion see Sakas (in press).[6]

3. The Language Domain

The following experiments were run on an artificial language domain created by the Computational Language Acquisition Group (CoLAG) at the City University of New York (CUNY) and consists of 3,072 languages, together with the parameterized grammars that generate them and syntactic derivations for all the word-order patterns that make up the languages in the domain. The 13 binary parameters that generate the CoLAG languages embody familiar syntactic differences between natural languages. They are listed in Table 1, with their values.

Parameter name	**0 / 1 value**	**English settings**
Subject Position	initial / final	0
Headedness in IP, NegP, VP, PP	final / initial	0
Headedness in CP	initial / final	0
Optional Topic	obligatory / optional	1
Null Subject	no null subject / null subject	0
Null Topic	no null topic / null topic	0

[6] This study assumes there no noise in the input stream. Noise can severely mitigate the STL's efficiency (cf., Crother, Fodor & Sakas, 2004) whereas the VL is likely to be more robust in the face of noise. The effect of noise on learnability for parameter setting models warrants further inquiry.

Wh-Movement	no Wh-movement / Wh-movement	1
Preposition Stranding	no stranding / stranding	1
Topic Marking	no marking / marking	0
VtoI Movement	no movement / movement	0
ItoC Movement	no movement / movement	0
Affix Hopping	no hopping / hopping	1
Q-Inversion: ItoC in questions	no inversion / inversion	1

Table 1: Parameters, their default values, and most English-like settings of the 13 parameters that were used to generate the 3,072 languages that make up the CUNY CoLAG language domain.

There are constraints on some parameters, e.g., languages are forbidden to have both Null Subject and Null Topic set to their positive values which is why there are 3,072 languages rather than 2^{13} (8,192) which would ensue without parameter constraints. A CoLAG grammar is a string of thirteen 1's and 0's corresponding to the thirteen parameters. CoLAG's version of 'English' would be: 1100011001000, which also the binary representation of the number 611.[7]

CoLAG languages consist of word-order patterns that encode sentences with tokens that denote the grammatical roles of words and complex phrases, e.g., subject (S), direct object (O1), indirect object (O2), main verb (V), auxiliary verb (Aux), adverb (Adv), preposition (P), etc. An example pattern is *S Aux V O1* which corresponds to the English sentence: *The little girl can make a paper airplane*. There are also tokens for topic and question markers for use when a grammar specifies overt topicalization or question marking. The languages consist of patterns with only a single sentential clause (degree-0) and since there is no (other) recursion the languages are finite; on average each language consists of 789 sentence patterns.

Each word-order pattern has two or more syntactic derivations in the form of fully specified X-bar style trees using GB-style phrase markers. A slash feature is borrowed from HPSG to encode movement as local dependencies. Other features include +NULL for non-audible tokens (e.g. S[+NULL] represents a null subject *pro*), +TOPIC to represent a topicalized token, +WH to denote wh-words (e.g., *who*, *what*, *where*), illocutionary (ILLOC) features, etc.

Figure 1, replicated from Sakas & Fodor (2012), depicts two different derivations for the CoLAG pattern *S Aux Verb Adv*.[8]

[7] Although we use 1's and 0's for notational convenience, how parameter values are linguistically manifested can vary from model to model (e.g., STL 'treelets' above) and between formalisms (e.g., GB vs. MP).

[8] Note that unlike Sakas & Fodor, we omit tense and illocutionary force features for simplicity of presentation.

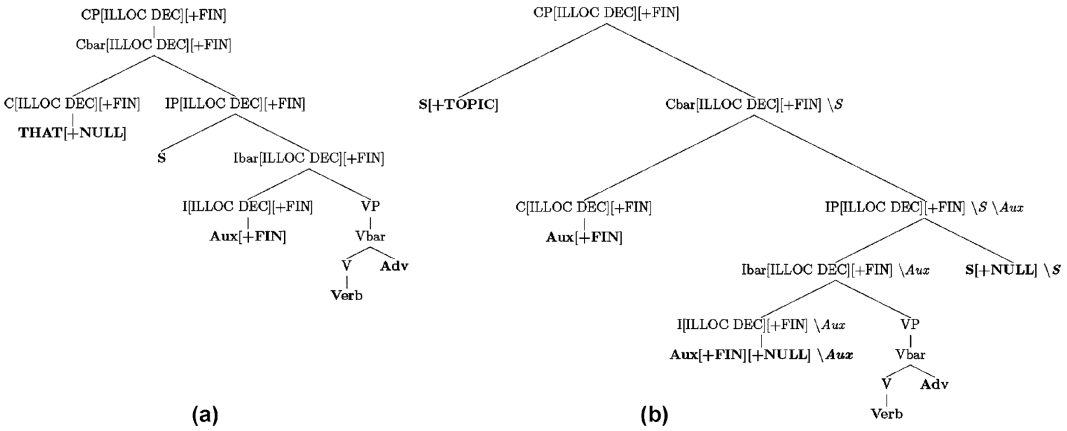

FIGURE 1 Two fully specified CoLAG tree structures for *S Aux Verb Adv*. These trees are from languages that differ with respect to two relevant parameters. The tree depicted in Figure 1a is generated by a grammar with the Subject-initial value of the Subject Position parameter, and the No Movement value of the ItoC Movement parameter. The tree in Figure 1b is generated by a grammar with the Subject-final value of the Subject Position parameter and the Obligatory Movement value of the ItoC Movement parameter.

This is an illustration of parametric ambiguity where different (relevant) parameter settings license the same sentence. The CoLAG domain has a substantial amount of ambiguity. As one measure, on average, each CoLAG sentence pattern can be licensed by 50 grammars. There are a total of 48,077 distinct sentence patterns in the CoLAG domain (across all languages), and 93,768 distinct syntactic trees. Figure 1 displays only two of the twelve distinct syntactic derivations of the sentence pattern *S Aux Verb Adv*.[9] As discussed below, parametric ambiguity is a significant obstacle all parameter setting models must overcome to succeed in acquiring the correct target grammar.

The CoLAG domain allows us to pursue a wide range of language learnability research which would be unmanageable in the domain of natural languages. The domain was designed to include many phenomena from the learnability literature as possible (e.g., subset-superset relations, non independence of parameters, null elements, etc.) while staying true to observed cross-linguistic data. That said, decisions had to be made in order to bring the domain into existence (e.g., GB formalism, thirteen parameters, etc.). All were carefully considered, based on: practicable concerns, embracing as many know learnability issues from developmental psycholinguistics as possible, and linguistic viability. However, even though the CoLAG domain is the largest domain of its kind with which one can systematically pose learnability questions that mirror those of natural language, other design decisions might very well change the shape of the learnability problems and solutions that arise. Extensive details of the CoLAG domain, and the 'supergrammar' that generated the domain can be found in Sakas & Fodor (2011, 2012).

---

[9] Ambiguity is often conflated with irrelevance. For example, if there is no preposition in a sentence, *s*, then the Preposition Stranding parameter is irrelevant—it can be set to either value and both corresponding grammars will license *s* with exactly the same tree derivation being constructed. This is quite different from the ambiguity depicted in Figure 1 where the tree structures are very different. The first measure given (50 grammars) conflates parametric irrelevance and ambiguity, the second (twelve derivations) doesn't. What constitutes the 'correct' measure depends on the learner being simulated.

The entire domain (and both articles) are available online at: http://www.colag.cs.hunter.cuny.edu/downloadables.html.

4. The Simulations.

Overview We ran three experiments. In all three experiments, we simulated three models of parameter setting, the Variational Learning (VL) model, The Structural Triggers Learner (STL), and the Triggering Learning Algorithm (TLA). They are described below.

Target grammar and input sentences As the target grammar, we used the CoLAG language that was generated by the parameter values most closely aligned with English (henceforth Co-English) and the learning models were tested on three different distributions of CoLAG sentence patterns. The first distribution was to supply all the sentence patterns from Co-English with a uniform distribution: there are 360 sentence patterns in Co-English. The second distribution was derived by mapping a parsed corpus of child-directed English (Pearl and Sprouse, 2013) to CoLAG tokens (*S*, *Verb*, *O1*, etc.) and using frequencies derived from the resulting corpus: there are 81 sentence patterns in Co-English that map to at least one sentence pattern in the parsed corpus of child-directed English. The third distribution used only the 81 sentence patterns, but with a uniform distribution untethered to the frequencies that exist in the parsed corpus.

Learning Trials All three learners are incremental, in the sense that the learning algorithms are applied to each independent input sentence, and not across multiple sentences (that is, there is no memory for previously encountered sentences). In all cases a single learning trial proceeded by exposing each model to sentence patterns from the target language (e.g., Co-English) one at a time. A trial ends when one of the following occurs: i) The learner successfully converges to the target grammar, i.e., all 13 parameters in the domain are set correctly for the target language or a grammar that generates a superset of the target language,[10] or ii) The learner encounters a maximum number of input sentences without setting all 13 parameters correctly, i.e., the learner fails to converge on the target before the threshold is reached. The average number of sentences consumed over 100 trials is reported as well as a percent succeed/fail rate. A failure was reported if the learner consumed 2 million input sentences before converging. The learning rate for the VL was fixed at .001.

Results Table 2 summarizes the results. The TLA does not consistently converge. The STL converges very rapidly. The VL model takes on average under half a million sentences to converge which is still well within the amount of data that children typically receive before the onset of combinatorial speech (Hart & Riesley 1995).[11]

---

[10] For mathematical and formal reasons, a target language and its supersets cannot be distinguished in the current setting. Also note that for the VL convergence was achieved when all the weights for the 13 parameters passed a threshold of .98 (the parameter considered set to its 1-value) or .02 (the parameter set to its 0-value).

[11] This is not entirely unproblematic. Although most of the parameters in the CoLAG domain are in fact setting quite early on as evidenced by the virtual absence of word order errors in child language

The TLA fails to take advantage of the structure of the CoLAG domain in terms of unambiguous triggers (or signatures) which is why it does not reliably converge (Gibson and Wexler, 1994, Berwick and Niyogi, 1996, Fodor, 1998a). That's not to say that there are not other domains in which the TLA might fare better. Sakas (2000), for example, presents a "strongly smooth" domain in which the TLA does perform well, however the domain is far less linguistically motivated than the current CoLAG domain.

Since the STL operates most efficiently when faced with unambiguous triggers we conjecture that the 81 CHILDES patterns contain a significantly higher percentage of unambiguous triggers than the entire set of 360 CoLAG-licensed Co-English patterns; convergence required only 20 sentence patterns for the CHILDES input sample, compared with 38 sentences for the full Co-English set of patterns given that the learner is equally as likely to see any sentence pattern at any given point of time. However, when the presentation of patterns mirrors that of child-directed speech the simulation results suggest that those unambiguous signatures are less likely to be encountered by the learner; on average the STL consumed 48 sentences to converge in this case.[12] Although the VL, as discussed above, requires signatures (or conditioned signatures) to converge efficiently, it is less sensitive to their proportion in the input stream due to the gradual nature of selecting hypotheses based on weights. In these simulations we used a relatively low learning rate (.001) which means that the weights were 'nudged' towards a 0- or 1-value in small increments leveling out convergence times.

This suggests that the structurally informative cues may be effectively combined with a gradual but robust probabilistic learning mechanism. In effect an STL learner supplied with activations as in Fodor (1998b) or Yang's (2002) weights that keeps track of successes in the face of ambiguity, would be a learning model that could take advantage of both the VL's smoothness and robustness and the STL's efficiency. Several versions can be envisioned, one might be: decode the input using the current bank of parametric treelets and reward those that do not need to be changed. When faced with a choice of parametric treelets that can be used by the parser to license the input, choose randomly but with probability conforming to the current weight (or activation) of that parameter.

It is important to note that for either the STL or VL learner, restricting the input sample to only those patterns found in CHILDES corpora did not prohibit the learners learners from converging. Clearly utterances encountered in actual child-directed speech are sufficient for a child to achieve competency of a full natural language. Likewise the CoLAG patterns occurring in child-directed speech are sufficient to 'trigger' the productivity realized in a full CoLAG language. Though future research needs to identify exactly what structural elements of the parameterized CoLAG domain led to these results

---

(Brown 1973), English-learning children famous undergo a protracted stage (Bloom 1970, Hyams 1986, Valian 1991) during which obligatory subjects--and occasionally objects--are omitted. This stage can be accounted for by the probabilistic use of a topic-drop grammar, which is gradually eliminated on the basis of expletive subjects as signature (e.g., *there is a cup on the table*; Yang 2002). Because the CoLAG domain does not include expletive subjects in the sentence patterns, this phenomenon cannot be captured.

[12] Unpacking this innocent enough conjecture will require significant analysis. Gross percentages are largely uninformative (e.g., more than 95% of declarative sentences set the headedness-in-IP-and-VP parameter in both input samples). Further study is required to pinpoint the proportions of triggers for specific parameters, and whether they are (or not) entangled with other parameters (Sakas & Fodor, 2012).

(see Footnote 12), they stand as the first computational demonstration of the potential power of parameters to compactly represent linguistic phenomenon that are robustly learnable.

	STL	%	VL	%	TLA	%
ALL Co-English Patterns	38	100	363,150	100	550	28
CHILDES patterns: Uniform distribution	20	100	464,551	100	486	93
CHILDES patterns: Empirical distribution	48	100	457,038	100	518	94

Table 2: Convergence rates in average number of sentence patterns consumed by each learner over 100 trials per simulation together with % success rates. Target grammar was Co-English from the CoLAG domain which generates 360 patterns. ALL used all 360 patterns, while the data derived from Pearl and Sprouse (2013) CHILDES corpora consist of 81 Co-English patterns.

5. Conclusion

Our study of parameter setting can be only regarded as preliminary. Indeed, it will always be a work in progress because the landscape of parameters is constantly changing as linguists get a better handle of the possibilities and constraints on syntactic variation. Nevertheless, we suggest that our results bode well with the parameter-based approach to language acquisition. At the minimum, we have shown that the parameters implemented in the CoLAG domain, which are drawn from extensive comparative research, appear to capture the range of syntactic variation in a compact and easily navigable way. In one sense, this should not be surprising: children do acquire an enormously complex grammatical system in a few short years, and the hypothesis space in which language learning operates must be favorable to learning. At the same time, our results can be regarded as a vindication of the parameter-based theory and its empirical reach: we now have a plausible answer of what such a favorable hypothesis space looks like. It will benefit not only the STL and VL models but all learning models that "modularize" the search for the target grammar along the dimensions specified by the parameters.

The promise of the parameters, in terms of both descriptive and explanatory adequacy, only raises questions about their places in a broad theory of language as a biological system. Surely there couldn't have been piecemeal evolution for each of the parameters under current study, and the success of the parameters must ultimately be attributed to deeper principles of language and related systems in human cognition—the goals of the minimalist program. A deeper understanding of how children learn, which may well employ mechanisms not specific to domain but shared across domains and species, will continue to shed light on the direction of linguistic research.

References

Berwick, Robert C. 1985. *The acquisition of syntactic knowledge.* Cambridge, MA: MIT Press.

Berwick, Robert C. & Partha Niyogi. 1996. Learning from triggers. *Linguistic Inquiry* 27, 605-622.

Bloom, Lois. 1970. *Language development: Form and function in emerging grammar.* Cambridge, MA: MIT Press.

Brown, Roger. 1973. *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht: Foris.

Crowther, Carrie, Janet Dean Fodor & William Gregory Sakas. 2004. Does ungrammatical input improve language learning? Paper presented at Architectures and Mechanisms for Language Processing Conference (AMLaP-2004), Universite de Provence.

Culicover, Peter W & Ray Jackendoff. 2005. *Simpler syntax.* Oxford University Press Oxford.

Dresher, B. Elan. 1999. Charting the learning path: Cues to parameter setting. *Linguistic Inquiry* 30, 27-67.

Fodor, Janet Dean. 1998a. Unambiguous triggers. *Linguistic Inquiry* 29, 1-36.

Fodor, Janet Dean. 1998b. Parsing to learn. *Journal of Psycholinguistic Research* 27, 339-374.

Fodor, Janet Dean & William Gregory Sakas. 2004. Evaluating models of parameter setting. In Alejna Brugos, Linnea Micciulla & Christine E Smith (eds.), *Proceedings of the 28th Annual Boston University Conference on Language Development (BUCLD 28)*, 1-27, Somerville, MA:Cascadilla Press.

Fodor, Janet Dean & William Gregory Sakas. 2005. The subset principle in syntax: Costs of compliance. *Journal of Linguistics* 41, 513-569.

Gibson, Edward & Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25, 407-454.

Gold, E. Mark. 1967. Language identification in the limit. *Information and Control* 10, 447-474.

Hart, Betty & Todd R Risley. 1995. *Meaningful differences in the everyday experience of young american children.* Baltimore, MD: Paul H Brookes Publishing.

Hyams, Nina. 1986. *Language acquisition and the theory of parameters.* Dordrecht: Reidel.

Kapur, S. (1994). Some applications of formal learning theory results to natural language acquisition. In B. Lust & G. Hermon (Eds.), *Syntactic Theory and First Language Acquisition: Cross-linguistic Perspectives. Volume 2: Binding, Dependencies, and Learnability* (pp. 491-508). Hillsdale, NJ: Lawrence Erlbaum.

Newmeyer, Frederick J. 2004. Typological evidence and universal grammar. *Studies in Language* 28, 527-548.

Niyogi, Partha. 2006. *The computational nature of language learning and evolution.* Cambridge, MA: MIT Press.

Pearl, Lisa & Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition* 20, 23--68. Data available on CHILDES.

Roeper, Thomas & Edwin Williams. 1987. *Parameter setting.* Dordrecht: Reidel.

Sakas, William Gregory. 2000. *Ambiguity and the computational feasibility of syntax acquisition*. New York, NY:City University of New York dissertation.

Sakas, William Gregory. 2016, to appear. Computational approaches to parameter setting in generative syntax. In J. Pater and W. Snyder J. Lidz (ed.), *Oxford handbook of developmental linguistics*. Oxford: Oxford University Press.

Sakas, William Gregory & Janet Dean Fodor. 2001. The structural triggers learner. In Stefano Bertolo (ed.), *Language acquisition and learnability*, 172-233. Cambridge: Cambridge University Press.

Sakas, William Gregory & Janet Dean Fodor. 2011. Generating colag languages using the 'supergrammar'. Online technical report: www.colag.cs.hunter.cuny.edu.

Valian, Virginia. 1991. Syntactic subjects in the early speech of american and italian children. *Cognition* 40, 21-81.

Valiant, Leslie G. 1984. A theory of the learnable. *Communications of the ACM* 27, 1134-1142.

Vapnik, Vladimir. 2000. *The nature of statistical learning theory.* Berlin: Springer.
Yang, Charles. 2002. *Knowledge and learning in natural language.* Oxford: Oxford University Press.

Yang, Charles. 2005. On productivity. *Yearbook of Language Variation*, 5, 333-370.

Yang, Charles. 2012. Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 3, 205-213.

Yang, Charles. 2016. *The price of linguistic productivity: How children learn to break rules of language.* Cambridge, MA: MIT Press.