1. Introduction

The ultimate problem in linguistics is to account for the acquisition of natural languages by children. Any reasonable theory must explain how they acquire a language and how they acquire it as rapidly as they do. Until recently most linguists were content to offer a promissory note on this problem, assuming instantaneous acquisition and then exploring the properties of a fully adult grammar. Now, however, there is a much greater concern among linguists as to what sorts of grammars are learnable, and how.

Stephen Pinker, in his article Formal Models of Language Acquisition (1979), surveys the literature to that point, and defines six conditions on theories of language acquisition:

1. The Learnability Condition is met if the theory can account for the fact that languages can be learned.

2. The Equipotentiality Condition is met if the theory does not succeed merely by being extremely narrow, forcing many things to be specified as innate where they can be learned. As a case in point, Pinker (1982, 1984) talks of children being "hard-wired" for all possible rules in a grammar, which then turn off permanently when presented with incompatible data. This theory satisfies Learnability but not Equipotentiality.

3. The Time Condition is met if the theory accounts for learning in the time children normally take for the acquisition of a grammar.
4. The Input Condition is met if the theory accounts for language learning with the typical input available to the child.

5. The Developmental Condition is met if the theory makes correct predictions about the child's capabilities during the course of acquisition.

6. The Cognitive Condition is met if the theory agrees with the known cognitive faculties of the child and the adult.

A formal theory of language acquisition, to be complete, must meet these conditions. There are no current models which can be rigidly shown to satisfy all six. Typically, models can succeed only with the Learnability, Time and Input conditions, if they succeed anywhere.

For our purposes, then, it is useful to define a formal model of language acquisition as a 4-tuple: $<\mathcal{G}, \mathcal{I}, \mathcal{C}, \Psi>$ where $\mathcal{G}$ is the set of grammars $G_0, G_1, \ldots, G_n$ posited by the language learner.

$\mathcal{I}$ is the set of input information $I_0, I_1, \ldots, I_n$ available to the language learner.

$\mathcal{C}$ is the criterion of success.

$\Psi$ is the procedure used to produce $G_i$ with input $I_i$.

Results from the psychological studies of children acquiring language serve as the data to refine each of these components within a model so that they will satisfy the Pinker conditions described above. No theory has advanced to this state yet,
and so, for the time being, mathematical results about formal languages are used to map out preliminary boundaries for each of these components, and to compare different models.

In section 2 I will discuss Gold's paper *Language Identification in the Limit* (1967), which is still the most important paper in the field. In the following sections I will discuss models of language learning based on three current flavors of generative grammar: the EST/GB model of Chomsky (Chomsky (1981, 1982); Wexler and Culicover (1980)); the Lexicalist-Functionalist model (Bresnan (1978, 1982)); and Generalized Phrase Structure Grammar (Gazdar (1982)). None of these models succeeds along all lines, but they approach the problem in different ways, illuminating problems shared by and specific to these models of grammars when they are applied as targets for language learners. Finally, I will look at some parsing results from computational linguistics and assess their relevance to language acquisition.

2. *Language Identification in the Limit*

Gold (1967) described a situation where a learner was given input one sentence at a time, and then was asked to guess a grammar to describe all input received to that point. The grammars were drawn from the Chomsky hierarchy of languages (Chomsky (1963)), and a class of languages were deemed to be learnable if a person could guess the correct grammar in a finite number of trials. Gold's special type of learnability is called learnability in the limit. A grammar is
learnable in the limit if the person eventually settles on a grammar, and that grammar is correct. That is, for sufficiently large $N$, $G_n = G_m = G$ for all $m, n > N$, where $G$ is the actual grammar. This shows that a criterion of success (C) can be a condition on the set of grammars (G).

Two different types of input are described: the most restrictive mode of presentation is text presentation, where the person is supplied only with examples of well-formed strings, one by one; the other method is informant presentation, in which both well-formed and ill-formed strings are presented with each one marked as grammatical or ungrammatical. Summarizing Gold's results:

<table>
<thead>
<tr>
<th>CLASS OF LANGUAGES</th>
<th>LEARNABLE?</th>
<th>CONTAINS NL's?</th>
</tr>
</thead>
<tbody>
<tr>
<td>recursive enumerable</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>recursive decidable</td>
<td>no</td>
<td>?</td>
</tr>
<tr>
<td>primitive recursive</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>context sensitive</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>context free</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>finite state</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>regular</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Peters and Ritchie (1973a) have shown that an Aspects-type grammar (Chomsky (1965)) provides a grammar for any recursive enumerable language. Thus Gold's result combined with this proof of the enormous generative capacity of the Aspects model shows that the Aspects-type grammars are unlearnable, even from an informant! This is an extremely powerful result
because it shows that the grammars most linguists had been working with must be severely constrained if they were to learnable at all.

It is not clear which is the most restrictive class of languages that can generate (in the weak or strong senses) all natural languages. Peters and Ritchie (1973a) argue that it is the context-sensitive class. Pullum and Gazdar (1982) argue for the context-free class and Reich (1969) argues for a finite-state description. Unfortunately Gold's results do not distinguish between these classes: they are all learnable from an informant and unlearnable from text. So, Gold's results cannot be used as an argument for any of these three classes.

Gold also showed that no learning process can be more effective under learnability in the limit than enumeration of the grammars in a class. So, learnability in the limit provides a precise criterion for success of a model and also provides an enumeration procedure which is an upper bound for all other procedures in efficiency. Thus it sets very firm limits on formal models of acquisition and provides a basis of judging the success of the models.

Gold has provided a rigid, well grounded definition for the Learnability Condition. If a model succeeds in providing a grammar that is learnable in the limit with some type of input (for example if it could be shown that children are in an effective informant presentation) then it is clear that the model meets the Learnability Condition.
3. Parametric and Degree-2 Learnability

Chomsky is the fore-most proponent of the innateness hypothesis regarding the acquisition of a grammar. He states (Chomsky (1965, 1981)) that, in his view, much of the structure of the grammar is present in the child at birth, that the child comes equipped with a Universal Grammar. In his more recent work (1981), he sees the grammar as being composed of sets of parameters set to the unmarked value at birth, and only changed when the child receives input that is evidence for the marked value. Given a large enough set of parameters this view can account easily for the Learnability Condition, however it would fail in other areas without amendment. It posits a great deal of structure, and thus does not easily satisfy the Equipotentiality Condition, and would fail on the Time, Input and Developmental Conditions. This theory predicts that a very small amount of data would serve to fix a parameter, implying that the child should learn constructions with the same number of parameters in roughly equal amounts of time, provided the child is given some examples to fix the parameters. This does not seem to be the case.

On the one hand, this theory would predict acquisition that is faster than that actually observed, as very small input sets are sufficient to fix the parameters. This does not account for the massive amount of input the child re-
ceives before mastering the question constructions. On
the other hand, some parameters, such as Subjacency, require
very complex input that the child may not receive. The
sentences necessary to set this parameter involve multiple
embeddings, and the child receives very few of these types
of sentences as input. Even if the child did receive
such input, it is not at all clear how they would discern
which were the relevant parameters to be fixed by the data.
So, this theory falls short of meeting the Developmental
Condition, for it fails to provide an account of intermediate
capability. In addition, sometimes it predicts too rapid
acquisition of structures, thus also failing on the Time
Condition. In short, the theory does not adequately dis-
tinguish between common and uncommon constructions in a
language, overgeneralizing a rapid learning procedure
throughout, thus not being realistic regarding development
and input. Finally, by positing that the language faculty
is very different from the other cognitive faculties, the
theory cannot be used as a model for other cognitive abilities,
a disappointing result. A theory which claimed similar
structures for all cognitive functions would be a more valued
theory, so this is a retreat from a strong claim before any
negative data is available to refute the strong claim.

Wexler and Culicover (1980) devote 650 pages to developing
a theory that will learn transformations based on restricted
input. Their system will learn transformations with sentences
involving no more than two embeddings. This is deemed
Degree-2 input. One result they derive from this is the Binary Principle, which amounts to Chomsky's Subjacency Condition. This may allay some criticisms of Chomsky's parameters until one examines this theory in detail, upon which major flaws become apparent. Using the abstract formal model: \(<G, I, C, \nu>\), their grammar component has a transformational syntax on a context-free base. Clearly, if transformations are unnecessary for the description of natural languages, then a theory accounting for the learning of transformations is also unnecessary. They beg the question of how the CF base is learned, showing only how a child could learn transformational rules when already equipped with a CF base. It is necessary for any formal model of acquisition employing phrase structure rules to show how these rules are learned by the child. Wexler and Culicover neglect this completely.

Wexler and Culicover also assume too much regarding the input available to the child. They use a modified text presentation, in which the child receives ordered pairs: \(<b, a>\) as input. The abscissa is the untransformed base generated sentence; the ordinate is the surface structure resulting from the application of the transformations. They licence this claim by citing MacNamara (1972), who showed that semantic information is available to the learner through non-linguistic means. However, MacNamara's real world result does not imply
that untransformed structures are available to the language learner in the system. Thus, this is an unrealistic idealization of the input available during language acquisition. Thus this model also fails to meet Pinker's Conditions on Formal Models of Language Acquisition.

4. Lexical-Functional Grammar

Joan Bresnan (1978, 1982) has proposed a theory of grammar more concerned with psychological reality than Chomsky's have been. In line with this aim she has recruited a number of people, Pinker among them, to study the different aspects of her formulation of a theory of grammar. Pinker (1982, 1984) has examined LFG with regard to learnability with largely favorable results compared with Chomsky's formulations. In LFG all c-structures are generated directly by a context-free grammar, so there are no transformations in the syntax to be learned. The other component of the grammar, the f-structure is a representation of the semantic content of the sentence. This formulation allows the child learning the grammar to have access to the semantic information (through non-linguistic means) and to apply this information to the acquisition of the grammar, in a plausible way. So, MacNamara's results are put to good and proper use here, whereas Wexler and Culicover put them to improper uses.

Transformations are "replaced" in LFG by lexical (redundancy) rules, effectively adding new subcategorizations for f-structures and c-structures from old ones. It also adds metavariables (++] for control and binding. Now, both of these additions
must be shown to be learnable by the child. Pinker shows that the problem of overgeneralization of lexical rules can be solved by only allowing lexical rules to be incorporated into the grammar for any particular lexical item when the child has actually heard the form. This vastly limits the productivity of the child's language but would preclude the overgeneralizations resulting in the generation of bad forms. He cites examples from the literature to support his claim of lessened productivity. His support for the learnability of LFG's is impressive and much more fully thought out than Wexler and Culicover's. He makes better use of actual child language data, and points out the weaknesses in the LFG theory himself. He takes pains to explain learning procedures for each component: c-structures using a small set of initial rules similar to the $\bar{X}$-theory (Jackendoff (1977)), f-structures using a semantic bootstrapping hypothesis (a set of general semantic heuristics), and lexical rules using a distributional analysis procedure.\textsuperscript{2,3}

Lexical-Functional Grammar has two general features which Pinker shows enhance its learnability: it restricts the syntax (for the most part) to a context-free grammar, and it employs a semantic component allowing the child to make use of the semantic information that is available. The only portion of LFG whose learnability is not clearly accounted for is the use of the metavariables. These can range in scope over great portions of c-structures, and it is not shown by Pinker or Bresnan how a child would receive the relevant data to learn the operations
of the metavariables, consistent with Pinker's Conditions on Formal Models.

5. Generalized Phrase Structure Grammar

Another recent variant of generative grammar has been proposed by Gerald Gazdar (1982) and others. Gazdar argues that a CF grammar is sufficient to generate all natural languages, and Pullum (1984), Pullum and Gazdar (1982) show that previous "proofs" that NLs are non-context-free are all fallacious. Gazdar proposes a CF grammar employing complex symbols, as in Harmen (1963), aided by conditions imposed on immediate constituent structures. This immediate constituent context-sensitive type of grammar has been shown by Peters and Ritchie (1973b) to be exactly the class of CF grammars, and so these conditions may be included in the grammar without increasing the generative power of the resulting system.

Gazdar is especially critical of attempts within TG to reduce the power of transformations while still retaining them. His view is that the "strongest way to constrain a component is to eliminate it." (1982: 132). He points out that saying such "restricted" TGs are less powerful is akin to saying that "Turing machines which employ narrow grey tape are less powerful than ones employing wide orange tape." (1982: 131).

Now, in addition to this removal of transformations, Gazdar views syntactic rules not as rewriting systems but as Node Admissability Conditions. The rule S → NP VP (in
Gazdar's notation $[\bar{\bar{N}} \bar{V}]$ does not replace the symbol string $\bar{V}$ $XSY$ by the symbol string $X$ NP VP $Y$ but rather states that two adjacent constituents $\bar{\bar{N}} \bar{V}$ may be combined as a constituent of type $\bar{V}$. (That is that $\bar{V}$ will dominate $\bar{\bar{N}}$ and $\bar{V}$ immediately and exclusively). Each of these NACs is associated with a CF semantic interpretation rule in a possible world intensional logic similar to Montague semantics. This allows the results obtained by MacNamara (1972) to be incorporated in a meaningful way into the theory. The child receives input pairs consisting of a sentence and its semantic interpretation (inferred from non-linguistic data), and uses these to build rules of the grammar, which are of form: $<a, [\ldots b \ldots ], a'>$ where the first $a$ component is the rule number, the second is the syntactic rule and the third is the associated semantic rule. Both the syntax and the semantics are structures built up from smaller constituents, starting from the lexical items. Now, this view of PS rules is preferable to the rewrite interpretations of Bresnan and Chomsky because it predicts certain facts about how children acquire syntactic complexity that the other theories do not. Children start at the one word level, gradually building up more complex structures. GPSG works the same way, building up phrase structures from lexical items. Thus GPSG satisfies in this way the Learnability, Developmental and Input Conditions in the learning of the grammar. The child begins with a word of a major category (N or V), perhaps incorporating Pinker's semantic bootstrapping heuristic that nouns are things and
verbs are actions. Then, using distributional analysis the learner figures out how these major categories are combined with other words to build more complex structures, for example that [A N], that is, an adjective followed by a noun, N constitutes a larger constituent with the semantic value of N with property A. Thus GPSG provides an explanation of how children learn the CF syntax that Wexler and Culicover just assumed was acquired, somehow, without explanation.

In order to capture generalizations about natural languages GPSG augments the basic set of "production" rules (NACs) with other, derived rules. This is accomplished in a principled way using metarules. Metarules act in the following manner: "if r is a rule of format R then F(r) is also a rule where F(r) is some function of r." (Gazdar (1982: 158)). Again, I believe that this is a more realistic theory than transformations with regard to the language learner. For example, the child will eventually acquire the metarule introducing conjoined structures. Now, GPSG allows the child to acquire these rules as PS rules to begin with, limiting their productivity to conjoining nodes for which the child has had positive evidence. Eventually, when enough similar conjunction PS rules for different node types have been learned, the learner generalizes these rules into the metarule introducing conjoined structures. At this time the productivity should increase dramatically, for new PS rules can be deduced in addition to being inferred. The child may now produce examples of conjoined structures which have not been presented to the
learner as input. The other important fact about metarules is that though they increase the effective number of PS rules in the system, they capture generalizations without increasing the generative power of the system at all. The system begins with a basic set of context-free rules, and no matter how many derived rules are induced and deduced, the system remains context-free.

Up to this point, GPSG and LFG seem to be about even in the learnability conditions. LFG could be modified to view PS rules as NACs, and then the same distributional analysis could be invoked to learn the syntactic rules in the same way. The two theories differ in their predictions regarding productivity: LFG predicts steady increases in productivity, whereas GPSG predicts initial increases as in LFG, but with a blossoming of productivity as metarules are formulated. There is another significant area in which LFG and GPSG differ: unbounded dependancies. In LFG unbounded dependancies are handled by the metavariables, which I pointed out are different from the rest of the system and pose problems for the learnability of LFGs. In GPSG unbounded dependancies are handled in the same way as all other derived rules: through metarules. There are metarules which introduce rules with slash categories. A slash category is of the form α/β where α and β are themselves categories. The interpretation is that a constituent α dominates all its usual possibilities except that it is missing a node of type β somewhere. In GPSG the most important slash categories are
those of type $a/NP$. Eventually the "hole" will be passed down constituent by constituent until the node $NP/NP$ is reached. At this point this node will dominate an empty category (in the sense of Chomsky (1982, 1981)). This is equivalent to Bresnan's rule $NP + e$. So, where Bresnan's use of metavariables is clearly not CF, Gazdar's method of accounting for unbounded dependencies keeps the system context-free.

Further, the child could learn the metarule introducing slash categories using the same distributional analysis used in learning the other metarules. The child has access to closely spaced pairs such as

(1) Which car do you want?

(2) Do you want the red car?

The child will learn that $want$ is a transitive verb (that is, in the class of $V$ which support $[V \, \bar{N}]$), and, armed with this knowledge, will realize that there is a hole in (1) that is not present in (2), leading to the analysis:

(1') Which car $[do$ you want $e]$?

The child will learn that the wh-phrase at the beginning of the question announces that the following constituent will contain a hole. Thus, in GPSG unbounded dependency facts are learnable through distributional analysis, preserving the continuity of $\bar{V}$, the learning procedure. Such is not the case for the formulations of Bresnan and Chomsky.
Thus the advantages of GPSG as a model of language learning are that it employs a CF syntax and a CF semantics, associated rule by rule (allowing MacNamara's findings to be incorporated); the syntax contains only PS rules for generating structures; generalizations regarding PS generated structures are captured through a single mechanism: metarules; this mechanism does not increase the generative power of the syntax, and explains productivity facts; one mechanism allows the learning of all components: distributional analysis. Thus GPSG makes several strong claims regarding a formal model of language acquisition:

\[ \mathcal{G} = \{ G_i \mid G_i = \{ \text{rules} \} \cup \{ \text{metarules} \} \} \]

\[ \mathcal{I} = \{ I_i \mid I_i = \langle \text{sentence}, \text{semantic interpretation}\rangle \} \]

\[ \mathcal{C} = \text{there exists } N \mid \text{for all } n, m \geq N \ G_n = G_m = G \text{ where } G \text{ is the adult (target) grammar} \]

\[ \mathcal{U} = \text{distributional analysis} \]

Thus, GPSG seems to be a genuine start towards a formal theory of language acquisition that will account for a wide variety of acquisition facts within a very restricted framework of innate mechanisms. GPSG is well along the way to satisfying the six conditions outlined in Pinker(1979), and so can be regarded as a fairly successful formal model, much more successful than other models of language acquisition based on formulations by Bresnan and Chomsky.
6. Parsing and Learnability

Computational Linguistics has recently started to take an interest in language acquisition, in order to design computer models of language acquisition. Space does not permit me to review these attempts here, but there are two programs in current research that have theoretical implications for the interaction of parsability and learnability.

Berwick (1980, 1984) and Berwick and Weinberg (1984) describe their LPARFISAL system based on the Marcus parser PARFISAL (Marcus, 1980). This is a stripped down version of the Marcus parser, with a minimal set of parsing rules, and a procedure allowing it to learn new ones. When the parser tries to parse an incoming sentence and fails, the system tries to build a new rule to allow the parse to continue, if the rule is successful, the parse resumes, and the rule is remembered. If the rule is not successful then the parser gives up and goes on to the next sentence. In this way the parser learns in a steady, non-instantaneous manner. This learning of PS rules is augmented with a procedure that creates transformational rules that seem to correspond to the conditions on transformations that Wexler and Culicover found, and that have been proposed by Chomsky. Berwick states that this shows "that parsability is a stronger constraint than learnability" (Berwick, 1984: 20). In this way, given the model Chomsky currently proposes, and Berwick's choice of innate conditions as a starting formulation of PARFISAL, PS rules and transformations are learned without any further constraints, showing that in this model constraints ensuring learnability
are implied by the constraints already imposed to ensure parsability.

J.D. Fodor (1984) has extended her treatment of Phrase Structure Parsing (1983) to address the overlap of constraints on learnability with those on parsability. She examines the need for degree-2 data for the language learner and concludes that if the child were acquiring a GPSG grammar that only access to data of degree-2 is necessary. No natural language has rules that apply one way to twice embedded S's and another way to once embedded S's. (Languages do distinguish between matrix and embedded S's in the application of certain rules.) A child can infer a PS grammar with this constraint simply by noting that multiply embedded constructions can occur, without reference to the internal structure of the lower clauses. Now learners can get by with degree-2 data without information about the lower clause constituents, this is termed degree-2 data.

In Phrase Structure Parsing, this condition that NL's not behave differently with regard to multiply embedded S's and rule application follows from the design of the parser, using a Minimal Attachment Strategy. Thus the condition that Wexler and Culicover need for their learnability (degree-2 data) is contained within the constraint that the parser imposes on NLs (degree-2'). Thus, the conditions necessary for easy learnability are subsumed by those necessary for easy parsability, and therefore parsability ensures learnability.
7. Conclusions

The preceding sections have summarized the current literature regarding formal models of language acquisition, specifically, the acquisition of syntax. In conclusion, I would like to point out that the results obtained by Fodor (1984) re-inforce the conclusions favoring the GPSG model. There are many known parsing algorithms for context-free languages, as all programming languages are context-free languages. Therefore, a CF model such as GPSG would enhance both parsability, and thus (by Fodor's result, and by Pinker's Conditions) would also enhance learnability.

The finding that children have access to both sentences and their semantic interpretation lends further support to a CF model. The syntax is learned through distributional analysis, and is CF; the semantics is learned through similar distributional analysis, and is also CF. Now, as Pullum (1984) points out, if NL strings with regard to both syntax and semantics turn out not to be context-free, this is still within the range of a GPSG, because the intersection of two CF languages is not necessarily a CF language. With these two components, and a parallel processing model (a realistic view of cognitive capability) both efficient parsing and efficient learning are possible. With this addition the GPSG also starts to fulfill the Cognitive Condition, making it an even more promising formal model of language acquisition.
Notes:

1 For arguments against this assumption see Matthews (1979).

2 Sydney Lamb apparently argues for a similar sort of learning procedure in his paper On the mechanization of syntactic analysis in the 1981 Conference on Machine Translation of Language and Applied Language Analysis, Volume II, pages 674-685. This is referred to in Gold (1967) but I was unable to obtain a copy.

3 Pinker (1979) argues against distributional analysis, giving the example:
   
   Hottentots must survive.
   Hottentots must fish.
   Hottentots eat fish.
   Hottentots eat rabbits.

   pointing out:
   
   "Hottentots eat survive.
   Hottentots must rabbits."

   While this is true, a slightly more sophisticated distributional analysis will yield good results: given a word formation rule in English N +φ => V and V +φ# => N, this predicts the forms above ungrammatical, as they should be, and also predicts:

   Hottentots must rabbit.

   Which I judge to be grammatical, on the basis that I have heard rabbit used as a verb meaning "to run away in cowardice".

4 The relevant metarule is [...] + [...β/γ...].
   
   \[ α \rightarrow α/γ \]
References:


