Cognition, 7 (1979) 217–283
©Elsevier Sequoia S.A., Lausanne – Printed in the Netherlands

Formal models of language learning*

STEVEN PINKER**
Harvard University

Abstract

Research is reviewed that addresses itself to human language learning by developing precise, mechanistic models that are capable in principle of acquiring languages on the basis of exposure to linguistic data. Such research includes theorems on language learnability from mathematical linguistics, computer models of language acquisition from cognitive simulation and artificial intelligence, and models of transformational grammar acquisition from theoretical linguistics. It is argued that such research bears strongly on major issues in developmental psycholinguistics, in particular, nativism and empiricism, the role of semantics and pragmatics in language learning, cognitive development, and the importance of the simplified speech addressed to children.

I. Introduction

How children learn to speak is one of the most important problems in the cognitive sciences, a problem both inherently interesting and scientifically promising. It is *interesting* because it is a species of the puzzle of induction: how humans are capable of forming valid generalizations on the basis of a finite number of observations. In this case, the generalizations are those that allow one to speak and understand the language of one's community, and are based on a finite amount of speech heard in the first few years of life. And language acquisition can claim to be a particularly *promising* example of this

*I am grateful to John Anderson, Roger Brown, Michael Cohen, Martha Danly, Jill de Villiers, Nancy Etcoff, Kenji Hakuta, Reid Hastie, Stephen Kosslyn, Peter Kugel, John Macnamara, Robert Matthews, Laurence Miller, Dan Slobin, and an anonymous reviewer for their helpful comments on earlier drafts of this paper. Preparation of this paper was supported in part by funds from the Department of Psychology and Social Relations, Harvard University; the author was supported by NRC and NSERC Canada Postgraduate Scholarships and by a Frank Knox Memorial Fellowship.

**Reprints may be obtained from the author, who is now at the Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

puzzle, promising to the extent that empirical constraints on theory construction promote scientific progress in a given domain. This is because any plausible theory of language learning will have to meet an unusually rich set of empirical conditions. The theory will have to account for the fact that all normal children succeed at learning language, and will have to be consistent with our knowledge of what language is and of which stages the child passes through in learning it.

It is instructive to spell out these conditions one by one and examine the progress that has been made in meeting them. First, since all normal children learn the language of their community, a viable theory will have to posit mechanisms powerful enough to acquire a natural language. This criterion is doubly stringent: though the rules of language are beyond doubt highly intricate and abstract, children uniformly *succeed* at learning them nonetheless, unlike chess, calculus, and other complex cognitive skills. Let us say that a theory that can account for the fact that languages can be learned in the first place has met the *Learnability Condition.* Second, the theory should not account for the child's success by positing mechanisms narrowly adapted to the acquisition of a particular language. For example, a theory positing an innate grammar for English would fail to meet this criterion, which can be called the *Equipotentiality Condition.* Third, the mechanisms of a viable theory must allow the child to learn his language within the time span normally taken by children, which is in the order of three years for the basic components of language skill. Fourth, the mechanisms must not require as input types of information or amounts of information that are unavailable to the child. Let us call these the *Time* and *Input Conditions*, respectively. Fifth, the theory should make predictions about the intermediate stages of acquisition that agree with empirical findings in the study of child language. Sixth, the mechanisms described by the theory should not be wildly inconsistent with what is known about the cognitive faculties of the child, such as the perceptual discriminations he can make, his conceptual abilities, his memory, attention, and so forth. These can be called the *Developmental* and *Cognitive Conditions,* respectively.

It should come as no surprise that no current theory of language learning satisfies, or even addresses itself to, all six conditions. Research in psychology has by and large focused on the last three, the Input, Developmental, and Cognitive Conditions, with much of the research directed toward further specifying or articulating the conditions themselves. For example, there has been research on the nature of the speech available to children learning language (see Snow and Ferguson, 1977), on the nature of children's early word combinations (e.g., Braine, 1963), and on similarities between linguistic and cognitive abilities at various ages (e.g., Sinclair-de Zwart, 1969). Less often, there have been attempts to construct theoretical accounts for one or more of such findings, such as the usefulness of parental speech to children (e.g., Newport, Gleitman, and Gleitman, 1977), the reasons that words are put together the way they are in the first sentences (e.g., Brown, 1973; Schlesinger, 1971), and the ways that cognitive development interacts with linguistic development (e.g., Slobin, 1973). Research in linguistics that has addressed itself to language learning at all has articulated the Equipotentiality Condition, trying to distinguish the kinds of properties that are universal from those that are found only in particular languages (e.g., Chomsky, 1965, 1973).

In contrast, the attempts to account for the acquisition of language itself (the Learnability Condition) have been disappointingly vague. Language Acquisition has been attributed to everything from "innate schematisms" to "general multipurpose learning strategies"; it has been described as a mere by-product of cognitive development, of perceptual development, of motor development, or of social development; it has been said to draw on "input regularities", "semantic relations", "perceived intentions", "formal causality", "pragmatic knowledge", "action schemas", and so on. Whether the mechanisms implicated by a particular theory are adequate to the task of learning human languages is usually left unanswered.

There are, however, several bodies of research that address themselves to the Learnability criterion. These theories try to specify which learning mechanisms will succeed in which ways, for which types of languages, and with which types of input. A body of research called *Grammatical Induction,* which has grown out of mathematical linguistics and the theory of computation, treats languages as formal objects and tries to prove theorems about when it is possible, in principle, to learn a language on the basis of a set of sentences of the language. A second body of research, which has grown out of artificial intelligence and cognitive simulation, consists of attempts to program computers to acquire languages and/or to simulate human language acquisition. In a third research effort, which has grown out of transformational linguistics, a learning model capable of acquiring a certain class of transformational grammars has been described. However, these bodies of research are seldom cited in the psychological literature, and researchers in developmental psycholinguistics for the most part do not seem to be familiar with them. The present paper is an attempt to remedy this situation. I will try to give a critical review of these formal models of language acquisition, focusing on their relevance to human language learning.

There are two reasons why formal models of language learning are likely to contribute to our understanding of how children learn to speak, even if none of the models I will discuss satisfies all of our six criteria. First of all, a theory that is powerful enough to account for the *fact* of language acquisition may be a more promising first approximation of an ultimately viable theory than one that is able to describe the *course* of language acquisition, which has been the traditional focus of developmental psycholinguistics. As the reader shall see, the Learnability criterion is extraordinarily stringent, and it becomes quite obvious when a theory cannot pass it. On the other hand, theories concerning the mechanisms responsible for child language per se are notoriously underdetermined by the child's observable linguistic behavior. This is because the child's knowledge, motivation, memory, and perceptual, motor, and social skills are developing at the same time that he is learning the language of his community.

The second potential benefit of formal models is the explicitness that they force on the theorist, which in turn can clarify many conceptual and substantive issues that have preoccupied the field. Despite over a decade and a half of vigorous debates, we still do not know that sort of a priori knowledge, if any, is necessary to learn a natural language; nor whether different sorts of input to a language learner can make his task easy or difficult, possible or impossible; nor how semantic information affects the learning of the syntax of a language. In part this is because we know so little about the mechanisms of language learning, and so do not know how to translate vague terms such as "semantic information" into the information structures that play a causal role in the acquisition process. Developing explicit, mechanistic theories of language learning may be the only way that these issues can be stated clearly enough to evaluate. It seems to be the consensus in other areas of cognitive psychology that mechanistic theories have engendered enormous conceptual advances in the understanding of mental faculties, such as long-term memory (Anderson and Bower, 1973), visual imagery (Kosslyn and Schwartz, 1977), and problem solving (Newell and Simon, 1973).

The rest of the paper is organized into eight sections. In Section II, I will introduce the vocabulary and concepts of mathematical linguistics, which serve as the foundation for research on language learnability. Sections III and IV present E. Gold's seminal theorems on language learnability, and the subsequent research they inspired. Section V describes the so-called "heuristic" language learning models, several of which have been implemented as computer simulations of human language acquisition. Sections VI and VII discuss the rationale for the "semantic" or "cognitive" approach to language learning, focusing on John R. Anderson's computer simulation of a semantics-based learner. Section VIII describes a model developed by Henry Hamburger, Kenneth Wexler, and Peter Culicover that is capable of learning transformational grammars for languages. Finally, in Section IX, I discuss the implications of this research for developmental psycholinguistics.

II. Formal Models of Language

In this section I define the elementary concepts of mathematical linguistics found in discussions of language learnability. More thorough accounts can be found in Gross (1972) and in Hopcroft and Ullman (1969).

Languages and Grammars

To describe a language in mathematical terms, one begins with a finite set of *symbols*, or a *vocabulary*. In the case of English, the symbols would be English words or morphemes. Any finite sequence of these symbols is called a *string*, and any finite or infinite collection of strings is called a *language*. Those strings in the language are called *sentences;* the strings not in the language are called *non-sentences*.

Languages with a finite number of sentences can be exhaustively described simply by listing the sentences. However, it is a celebrated observation that natural and computer languages are infinite, even though they are used by beings with finite memory. Therefore the languages must have some finite characterization, such as a recipe or program for specifying which sentences are in a given language. A *grammar*, a set of rules that generates all the sentences in a language, but no non-sentences, is one such characterization. Any language that can be generated by a set of rules (that is, any language that is not completely arbitrary) is called a *recursively enumerable* language.

A grammar has four parts. First of all, there is the vocabulary, which will now be called the *terminal vocabulary* to distinguish it from the second component of the grammar, called the *auxiliary vocabulary*. The auxiliary vocabulary consists of another finite set of symbols, which may not appear in sentences themselves, but which may act as stand-ins for groups of symbols, such as the English"noun", "verb", and "prepositional phrase". The third component of the grammar is the finite set of *rewrite rules*, each of which replaces one sequence of symbols, whenever it occurs, by another sequence. For example, one rewrite rule in the grammar for English replaces the symbol "noun phrase" by the symbols "article noun"; another replaces the symbol "verb" by the symbol "grow". Finally, there is a special symbol, called the *start symbol*, usually denoted *S*, which initiates the sequence of rule operations that generate a sentence. If one of the rewrite rules can rewrite the "S" as another string of symbols it does so; then if any rule can replace part or all of that new string by yet another string, it follows suit. This procedure continues, one rule taking over from where another left off, until no auxiliary symbols remain, at which point a sentence has been generated. The language is simply the set of all strings that can be generated in this way.

Classes of Languages

There is a natural way to subdivide grammars and the languages they generate into classes. First, the grammars of different sorts of languages make use of different types of rewrite rules. Second, these different types of languages require different sorts of computational machinery to produce or recognize their sentences, using various amounts of working memory and various ways of accessing it. Finally, the theorems one can prove about language and grammars tend to apply to entire classes of languages, delineated in these ways. In particular, theorems on language learnability refer to such classes, so I will discuss them briefly.

These classes fall into a hierarchy (sometimes called the *Chomsky hierarchy*), each class properly containing the languages in the classes below it. I have already mentioned the largest class, the recursively enumerable languages, those that have grammars that generate all their member sentences. However, not all of these languages have a *decision procedure,* that is, a means of determining whether or not a given string of symbols is a sentence in the language. Those that have decision procedures are called *decidable* or *recursive* languages. Unfortunately, there is no general way of knowing whether a recursively enumerable language will turn out to be decidable or not. However, there is a very large subset of the decidable languages, called the *primitive recursive* languages, whose decidability *is* known. It is possible to *enumerate* this class of languages, that is, there exists a finite procedure called a *grammar-grammar* capable of listing each grammar in the class, one at a time, without including any grammar not in the class. (It is not hard to see why this is impossible for the class of decidable languages: one can never be sure whether a given language is decidable or not.)

The primitive recursive languages can be further broken down by restricting the form of the rewrite rules that the grammars are permitted to use. *Context-sensitive* grammars contain rules that replace a single auxiliary symbol by a string of symbols whenever that symbol is flanked by certain neighboring symbols. *Context-free* grammars have rules that replace a single auxiliary symbol by a string of symbols regardless of where that symbol occurs. The rules of *finite state* grammars may replace a single auxiliary symbol only by another auxiliary symbol plus a terminal symbol; these auxiliary symbols are often called *states* in discussions of the corresponding sentence-producing machines. Finally, there are grammars that have no auxiliary symbols, and hence these grammars can generate only a finite number of strings altogether. Thus they are called *finite cardinality* grammars. This hierarchy is summarized in Table 1, which lists the classes of languages from most to least inclusive.

Table 1. *Classes of Languages*

Class	Learnable from an informant?	Learnable from a text?	Contains natural languages?
Recursively Enumerable	no	no	yes*
Decidable (Recursive)	no	no	?
Primitive Recursive	yes	no	?
Context-Sensitive	yes	no	?
Context-Free	yes	no	no
Finite State	yes	no	no
Finite Cardinality	yes	yes	no

*by assumption.

Natural Languages

Almost all theorems on language learnability, and much of the research on computer simulations of language learning, make reference to classes in the Chomsky hierarchy. However, unless we know where natural languages fall in the classification, it is obviously of little psychological interest. Clearly, natural languages are not of finite cardinality; one can always produce a new sentence by adding, say, "he insists that" to the beginning of an old sentence. It is also not very difficult to show that natural languages are not finite state: as Chomsky (1957) has demonstrated, finite state grammars cannot generate sentences with an arbitrary number of embeddings, which natural languages permit (e.g., "he works", "either he works or he plays", "if either he works or he plays, then he tires", "since if either he...", etc.). It is more difficult, though not impossible, to show that natural languages are not context-free (Gross, 1972; Postal, 1964). Unfortunately, it is not clear how much higher in the hierarchy one must go to accomodate natural languages. Chomsky and most other linguists (including his opponents of the "generative semantics" school) use *transformational* grammars of various sorts to describe natural languages. These grammars generate bracketed strings called *deep structures,* usually by means of a context-free grammar, and then, by means of rewrite rules called *transformations,* permute, delete, or copy elements of the deep structures to produce sentences. Since transformational grammars are constructed and evaluated by a variety of criteria, and not just by the ability to generate the sentences of a language, their place in the hierarchy is uncertain. Although the matter is by no means settled, Peters and Ritchie (1973) have persuasively argued that the species of transformational grammar necessary for generating natural languages can be placed in the context-sensitive class, as Chomsky conjectured earlier (1965, p. 61). Accordingly, in the sections following, I will treat the set of all existing and possible human languages as a subset of the context-sensitive class.

III. Grammatical Induction: Gold's Theorems

Language Learning as Grammatical Induction

Since people presumably do not consult an internal list of the sentences of their language when they speak, knowing a particular language corresponds to knowing a particular set of rules of some sort capable of producing and recognizing the sentences of that language. Therefore learning a language consists of inducing that set of rules, using the language behavior of the community as evidence of what the rules must be. In the paragraphs following I will treat such a set of rules as a grammar. This should not imply the belief that humans mentally execute rewrite rules one by one before uttering a sentence. Since every grammar can be translated into a left-to-right sentence producer or recognizer, "inducing a grammar" can be taken as shorthand for acquiring the ability to produce and recognize just those sentences that the grammar generates. The advantage of talking about the grammar is that it allows us to focus on the process by which a particular language is learned (i.e., as opposed to some other language), requiring no commitment as to the detailed nature of the production or comprehension process in general (i.e., the features common to producers or recognizers for *all* languages).

The most straightforward solution to this induction problem would be to find some algorithm that produces a grammar for a language given a sample of its sentences, and then to attribute some version of this algorithm to the child. This would also be the most *general* conceivable solution. It would not be necessary to attribute to the child any a priori knowledge about the particular type of language that he is to learn (except perhaps that it falls into one of the classes in the Chomsky hierarchy, which could correspond to some putative memory or processing limitation). We would not even have to attribute to the child a special language acquisition faculty. Since a grammar is simply one way of talking about a computational procedure or set of rules, an algorithm that could produce a grammar for a language from a sample of sentences could also presumably produce a set of rules for a different sort of data (appropriately encoded), such as rules that correctly classify the exemplars and non-exemplars in a laboratory concept attainment task. In that case it could be argued that the child learned language via a general induction procedure, one that simply "captured regularity" in the form of computational rules from the environment.

Unfortunately, the algorithm that we need does not exist. An elementary theorem of mathematical linguistics states that there are an infinite number of different grammars that can generate any finite set of strings. Each grammar will make different predictions about the strings not in the set. Consider the sample consisting of the single sentence "the dog barks". It could have been taken from the language consisting of: 1) all three-word strings; 2) all article-noun-verb sequences; 3) all sentences with a noun phrase; 4) that sentence alone; 5) that sentence plus all those in the July 4, 1976 edition of the New York Times; as well as 6) all English sentences. When the sample consists of more than one sentence, the class of possible languages is reduced but is still infinitely large, as long as the number of sentences in the sample is finite. Therefore it is impossible for *any* learner to observe a finite sample of sentences of a language and always produce a correct grammar for the language.

Language Identification in the Limit

Gold (1967) solved this problem with a paradigm he called *language identification in the limit.* The paradigm works as follows: time is divided into discrete trials with a definite starting point. The teacher or environment "chooses" a language (called the *target language*) from a predetermined class in the hierarchy. At each trial, the learner has access to a single string. In one version of the paradigm, the learner has access sooner or later to all the sentences in the language. This sample can be called a *text*, or *positive information presentation.* Alternately, the learner can have access to both grammatical sentences and ungrammatical strings, each appropriately labelled. Because this is equivalent to allowing the learner to receive feedback from a native informant as to whether or not a given string is an acceptable sentence, it can be called *informant* or *complete information presentation*. Each time the learner views a string, he must guess what the target grammar is. This process continues forever, with the learner allowed to change his mind at any time. If, after a finite amount of time, the learner always guesses the same grammar, and if that grammar correctly generates the target language, he is said to have *identified the language in the limit.* Is is noteworthy that by this definition the learner can never know when or even whether he has succeeded. This is because he can never be sure that future strings will not force him to change his mind.

Gold, in effect, asked: How well can a completely general learner do in this situation? That is, are there any classes of languages in the hierarchy whose members can all be identified in the limit? He was able to prove that language learnability depends on the information available: if both sentences and non-sentences are available to a learner (informant presentation), the class of primitive recursive languages, and all its subclasses (which include the natural languages) are learnable. But if only sentences are available (text presentation), *no* class of languages other than the finite cardinality languages is learnable.

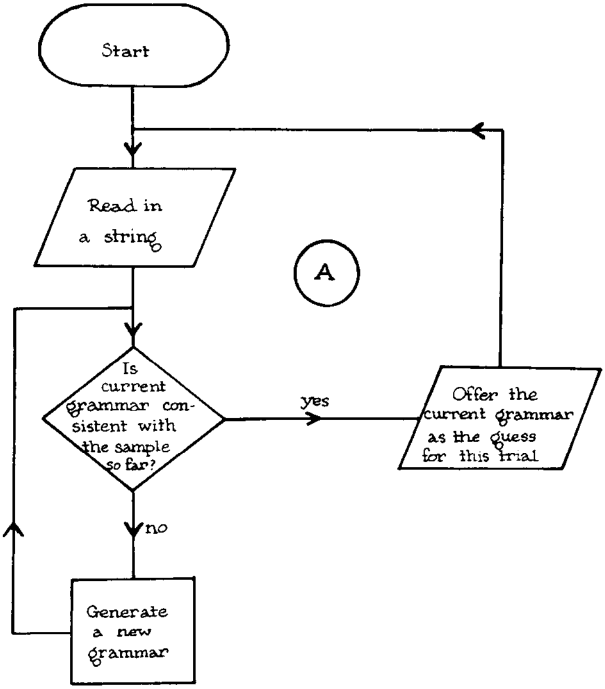
The proofs of these theorems are straightforward. The learner can use a maximally general strategy: he enumerates every grammar of the class, one at a time, rejecting one grammar and moving on to the next whenever the grammar is inconsistent with any of the sample strings (see Figure 1). With informant presentation, any incorrect grammar will eventually be rejected when it is unable to generate a sentence in the language, or when it generates a string that the informant indicates is not in the language. Since the correct grammar, whatever it is, has a definite position in the enumeration of grammars, it will be hypothesized after a finite amount of time and there will never again be any reason to change the hypothesis. The class of primitive recursive languages is the highest learnable class because it is the highest class whose languages are decidable, and whose grammars and decision procedures can be enumerated, both necessary properties for the procedure to work.

The situation is different under text presentation. Here, finite cardinality languages are trivially learnable – the learner can simply guess that the language is the set of sentences that have appeared in the sample so far, and when every sentence in the language has appeared at least once, the learner will be correct. But say the class contains all finite languages and at least one infinite language (as do classes higher than finite cardinality). If the learner guesses that the language is just the set of sentences in the sample, then when the target language is infinite the learner will have to change his mind an infinite number of times. But if the learner guesses only infinite languages, then when the target language is finite he will guess an incorrect language and will never be forced to change his mind. If non-sentences were also available, any overgeneral grammar would have been rejected when a sentence that it was capable of generating appeared, marked as a non-sentence. As Gold put it, "the problem with text is that if you guess too large a language, the sample will never tell you you're wrong".

Implication of Gold's theorems

Do children learn from a text or an informant? What evidence we have strongly suggests that children are not usually corrected when they speak ungrammatically, and when they are corrected they take little notice (Braine, 1971; Brown and Hanlon, 1970; McNeill, 1966). Nor does the child seem to have access to more indirect evidence about what is not a sentence. Brown and Hanlon (1970) were unable to discern any differences in how parents responded to the grammatical versus the ungrammatical sentences of their children. Thus the child seems to be in a text situation, in which Gold's

Figure 1. *A flowchart for Gold's enumeration procedure. Note that there is no "stop" symbol; the learner samples strings and guesses grammars forever. If the learner at some point enters loop "A" and never leaves it, he has identified the language in the limit.*


Start
Read in a string
A
Is current grammar consistent with the sample so far?
yes
no
Offer the current grammar as the guess for this trial
Generate a new grammar

learner must fail. However, all other models must fail in this situation as well – there can be no learning procedure more powerful than the one that enumerates all the grammars in a class.

An even more depressing result is the astronomical amount of time that the learning of most languages would take. The enumeration procedure, which gives the learner maximum generality, exacts its price: the learner must test astronomically large numbers of grammars before he is likely to hit upon the correct one. For example, in considering all the finite state grammars that use seven terminal symbols and seven auxiliary symbols (states), which the learner must do before going on to more complex grammars, he must test over a googol (10^{100}) candidates. The learner's predicament is reminiscent of Jorge Luis Borges's "librarians of Babel", who search a vast library containing books with all possible combinations of alphabetic characters for the book that clarifies the basic mysteries of humanity. Nevertheless, Gold has proved that *no* general procedure is uniformly faster than his learner's enumeration procedure. This is a consequence of the fact that an infinite number of grammars is consistent with any finite sample. Imagine a rival procedure of any sort that correctly guessed a certain language at an earlier trial than did the enumeration procedure. In that case the enumeration procedure must have guessed a different language at that point. But the sample of sentences up to that point could have been produced by many different grammars, including the one that the enumeration procedure mistakenly guessed. If the target language had happened to be that other language, then at that time the enumeration procedure would have been correct, and its rival incorrect. Therefore, for every language that a rival procedure identifies faster than the enumeration procedure, there is a language for which the reverse is true. A corollary is that every form of enumeration procedure (i.e., every order of enumeration) is, on the whole, equivalent in speed to every other one.

Gold's model can be seen as an attempt to construct some model, any model, that can meet the Learnability Condition. But Gold has shown that even if a model is unhindered by psychological considerations (i.e., the Developmental, Cognitive, and Time Conditions), learnability cannot be established (that is, unless one flagrantly violates the Input Condition by requiring that the learner receive negative information). What's more, no model can do better than Gold's, whether or not it is designed to model the child. However, since children presumably do have a procedure whereby they learn the language of their community, there must be some feature of Gold's learning paradigm itself that precludes learnability, such as the criterion for success or access to information. In Section IV, I will review research inspired by Gold's theorems that tries to establish under what conditions language learnability from a sample of sentences is possible.

IV. Grammatical Induction: Other Results

Grammatical Induction from a Text

This section will describe four ways in which languages can be learned from samples of sentences. One can either restrict the order of presentation of the sample sentences, relax the success criterion, define a statistical distribution over the sample sentences, or constrain the learner's hypotheses.

Order of sentence presentation

In Section III it was assumed that the sample strings could be presented to the learner in any order whatsoever. Gold (1967) proved that if it can be known that the sample sentences are ordered in some way as a function of time, then all recursively enumerable languages are learnable from a positive sample. Specifically, it is assumed that the "teacher" selects the sentence to be presented at time *t* by consulting a *primitive recursive function* that accepts a value of *t* as input and produces a sentence as output. Primitive recursive functions in this case refer to primitive recursive grammars that associate each sentence in the language with a unique natural number. Like primitive recursive grammars, they can be enumerated and tested, and the learner merely has to identify in the limit which function the teacher is using, in the same way that the learner discussed in Section III (and illustrated in Figure 1) identified primitive recursive grammars. This is sufficient to generate the sentences in the target language (although not necessarily sufficient to recognize them). Although it is hard to believe that every sentence the child hears is uniquely determined by the time that has elapsed since the onset of learning, we shall see in Section VI how a similar learning procedure allows the child to profit from semantic information.

Another useful type of sequencing is called *effective approximate ordering* (Feldman, 1972). Suppose that there was a point in time by which every grammatical sentence of a given length or less had appeared in the sample. Suppose further that the learner can calculate, for any length of sentence, what that time is. Then, at that point, the learner can compute all the strings of that length or less that are *not* in the language, namely, the strings that have not yet appeared. This is equivalent to having access to non-sentences; thus learning can occur. Although it is generally true that children are exposed to longer and longer sentences as language learning proceeds (see Snow

[redacted]

systematic changes in the speech directed to the developing child (see Snow and Ferguson, 1977) contain information that is useful to the task of inducing a grammar, as Clark (1973) and Levelt (1973) have suggested. For example, if it were true that sentences early in the sample were always generated by fewer rules or needed fewer derivational steps than sentences later in the sample, perhaps a learner could reject any candidate grammar that used more rules or steps for the earlier sentences than for the later ones. However, the attempts to discern such an ordering in parental speech have been disappointing (see Newport *et al.*, 1977) and it remains to be seen whether the speech directed to the child is sufficiently well-ordered with respect to this or any other syntactic dimension for an order-exploiting strategy to be effective. I will discuss this issue in greater depth in Section IX.

Relaxing the success criterion

Perhaps the learner should not be required to identify the target language exactly. We can, for example, simply demand that the learner *approach* the target language, defining approachability as follows (Biermann and Feldman, 1972; Feldman, 1972): 1) Every sentence in the sample is eventually included in the language guessed by the learner; 2) any incorrect grammar will at some point be permanently rejected; and 3) the correct grammar will be guessed an infinite number of times (this last condition defining *strong approachability*). The difference between strong approachability and identifiability is that, in the former case, we do not require the learner to stick to the correct grammar once he has guessed it. Feldman has shown that the class of primitive recursive languages is approachable in the limit from a sample of sentences.

The success criterion can also be weakened so as to allow the learner to identify a language that is an *approximation* of the target language. Wharton (1974) proposes a way to define a *metric* on the set of languages that use a given terminal vocabulary, which would allow one to measure the degree of similarity between any two languages. What happens, then, if the learner is required to identify any language whatsoever that is of a given degree of similarity to the target language? Wharton shows that a learner can approximate any primitive recursive language to any degree of accuracy using only a text. Furthermore, there is always a degree of accuracy that can be imposed on the learner that will have the effect of making him choose the target language exactly. However, there is no way of knowing how high that level of accuracy must be (if there were, Gold's theorem would be false). Since it is unlikely that the child ever duplicates exactly the language of his community, Wharton and Feldman have shown that a Gold-type learner *can* meet the Learnability condition if it is suitably redefined.

There is a third way that we can relax the success criterion. Instead of asking for the *only* grammar that fits the sample, we can ask for the *simplest* grammar from among the infinity of candidates. Feldman (1972) defines the *complexity* of a grammar, given a sample, as a joint function (say, the sum) of the *intrinsic complexity* of the grammar (say, the number of rewrite rules) and the *derivational complexity* of the grammar with respect to the sample (say, the average number of steps needed to generate the sample sentences). He then describes a procedure which enumerates grammars in order of increasing intrinsic complexity, thereby finding the simplest grammar that is consistent with a positive sample. However it is important to point out that such a procedure will *not* identify or even strongly approach the target language when it considers larger and larger samples. It is easy to see why not. There is a grammar of finite complexity that will generate every possible string from a given vocabulary. If the target language is more complex than this *universal grammar*, it will never even be considered, because the universal grammar will always be consistent with the text and occurs earlier in the enumeration than the target grammar (Gold, 1967). Thus equipping the child with Occam's Razor will not help him learn languages.

Bayesian grammar induction

If a grammar specifies the probabilities with which its rules are to be used, it is called a *stochastic* grammar, and it will generate a sample of sentences with a predictable statistical distribution. This constitutes an additional source of information that a learner can exploit in attempting to identify a language.

Horning (1969) considers grammars whose rewrite rules are applied with fixed probabilities. It is possible to calculate the *probability of a sentence given a grammar* by multiplying together the probabilities of the rewrite rules used to generate the sentence. One can calculate the *probability of a sample of sentences* with respect to the grammar in the same way. In Horning's paradigm, the learner also knows the *a priori probability* that any grammar will have been selected as the target grammar. The learner enumerates grammars in approximate order of decreasing a priori probability, and calculates the probability of the sample with respect to each grammar. He then can use the equivalent of Bayes's Theorem to determine the *a posteriori probability* of a grammar given the sample. The learner always guesses the grammar with the highest a posteriori probability. Horning shows how an algorithm of this sort can converge on the most probable correct grammar for any text.

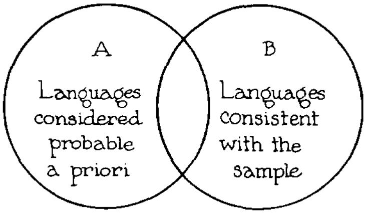
Constraining the hypothesis space

In its use of a priori knowledge concerning the likelihood that certain types of languages will be faced, Horning's procedure is like a stochastic version of Chomsky's (1965) abstract description of a language acquisition device. Chomsky, citing the infinity of grammars consistent with any finite sample, proposes that there is a *weighting function* that represents the child's selection of hypothesis grammars in the face of a finite sample. The weighting function assigns a "scattered" distribution of probabilities to grammars, so that the candidate grammars that incorporate the basic properties of natural languages are assigned high values, while those (equally correct) grammars that are not of this form are assigned extremely low or zero values. In weighting grammars in this way, the child is making assumptions about the probability that he will be faced with a particular type of language, namely, a natural language. If his weighting function is so constructed that only one highly-weighted grammar will be consistent with the sample once it has grown to a certain size, then learnability from a text is possible. To take an artificial example, if the child gave high values only to a set of languages with completely disjoint vocabularies (e.g., Hindi, Yiddish, Swahili, etc.), then even a single sentence would be sufficient evidence to learn a language. However, in Gold's paradigm, a learner that assigned weights of zero to some languages would fail to learn those languages should they be chosen as targets. But in the case of the child, this need not be a concern. We need only show how the child is able to learn human languages; it would not be surprising if the child was thereby rendered unable to learn various gerrymandered or exotic languages.

There are two points to be made about escaping Gold's conclusions by constraining the learner's hypothesis set. First, we lose the ability to talk about a general rule-inducing strategy constrained only by the computation-theoretic "lines of fracture" separating classes of languages. Instead, we are committed to at least a weak form of nativism, according to which "the child approaches the data with the presumption that they are drawn from a language of an antecedently well-defined type" (Chomsky, 1965, p. 27). Second, we are begging the question of whether the required weighting function exists, and what form it should take. It is not sufficient simply to constrain the learner's hypotheses, even severely. Consider Figure 2, a Venn diagram representing the set of languages assigned high a priori values (Circle A) and the set of languages that are consistent with the sample at a given point in the learning process (Circle B). To ensure learnability, the set of languages in the intersection between the two circles must shrink to a single member as more and more of the sample is considered. Circle B must not encompass Circle A completely, nor coincide with it, nor overlap with it to a large degree (a priori set too broad); nor can it be disjoint from it (a priori set too narrow). Specifying an a priori class of languages with these properties corresponds to the *explanatory adequacy* requirement in transformational linguistics. In Section VIII I shall examine an attempt to prove learnability in this way.

We have seen several ways to achieve learnability, within the constraint that only grammatical sentences be available to the learner. However, in

Figure 2. *Achieving learnability by constraining the learner's hypothesis set.*


A
Languages considered probable a priori
B
Languages consistent with the sample

severing one head of this hydra, we see that two more have grown in its place. The learning procedures discussed in this section still require astronomical amounts of time. They also proceed in an implausible manner, violating both the Developmental and the Cognitive criteria. First, children do not adopt and jettison grammars in one piece; they seem to add, replace, and modify individual rules (see Brown, 1973). Second, it is unreasonable to suppose that children can remember every sentence they have heard, which they must do to test a grammar against "the sample". In the next paragraphs I will review some proposals addressed to the Time Condition, and in Section V, research addressed more directly to the Developmental and Cognitive Conditions.

Reducing Learning Time

Efficient enumeration

The learners we have considered generate grammars rather blindly, by using a grammar-grammar that creates rules out of all possible combinations of symbols. This process will yield many grammars that can be shown to be undesirable even before they are tested against the sample. For example, grammars could be completely equivalent to other grammars except for the names of their auxiliary symbols; they could have some rules that grind to a halt without producing a sentence, and others that spin freely without affecting the sentence that the other rules produce; they could be redundant or ambiguous, or lack altogether a certain word known to appear in the language. Perhaps our estimate of the enormous time required by an enumeration procedure is artificially inflated by including various sorts of silly or bad grammars in the enumeration. Wharton (1977) has shown that if a learner had a "quality control inspector" that rejected these bad grammars before testing them against the sample, he could save a great deal of testing time. Furthermore, if the learner could reject not one but an entire set of grammars every time a single grammar failed a quality control test or was incompatible with the sample, he could save even more time, a second trick sometimes called *grammatical covering* (Biermann and Feldman, 1972; Horning, 1969; Wharton, 1977; Van der Mude and Walker, 1978). Horning and Wharton have implemented various enumeration techniques as computer programs in order to estimate their efficiency, and have found that these "quality control" and "covering" strategies are faster than blind enumeration by many orders of magnitude. Of course, there is no simple way to compare computation time in a digital computer with the time the brain would take to accomplish an analogous computation, but somehow, the performance of the efficient enumeration algorithms leaves little cause for optimism. For example, these techniques in one case allowed an IBM 360 computer to infer a finite state grammar with two auxiliary symbols and two terminal symbols after several minutes of computation. However natural languages have on the order of 10–100 auxiliary symbols, and in general the number of grammars using n auxiliary symbols grown as 2^{n^3}. Clearly, stronger medicine is needed.

Ordering by a priori probability

The use of an a priori probability metric over the space of hypothesis grammars, which allowed Horning's procedure to learn a language without an informant, also reduces the average time needed for identification. Since Horning's learner must enumerate grammars in approximate order of decreasing a priori probability, the grammars most likely to have been chosen as targets are also the ones first hypothesized. Thus countless unlikely grammars need never be considered. Similarly, if the learner could enumerate the "natural grammars" before the "unnatural" ones, he would learn more quickly than he would if the enumeration order was arbitrary. Unfortunately, still not quickly enough. Despite its approximate ordering by a priori probability, Horning's procedure requires vast amounts of computation in learning even the simplest grammars; as he puts it, "although the enumeration procedure... is formally optimal, its Achilles's heal is efficiency". Similarly, the set of natural languages is presumably enormous, and more or less equiprobable as far as the neonate is concerned; thus even enumerating only the natural languages would not be a shortcut to learning. In general, the problem of learning by enumeration within a reasonable time bound is likely to be intractable. In the following section I describe the alternative to enumeration procedures.

V. Heuristic Grammar Construction

Algorithms and Heuristics for Language Learning

Like many other computational problems, language learning can be attempted by *algorithmic* or *heuristic* techniques (see Newell and Simon, 1973). The enumerative procedures we have been discussing are algorithmic in that they guarantee a solution in those cases where one exists.[1] Unfortunately they are also prohibitively time-consuming and wildly implausible as models of children. Heuristic language learning procedures, on the other hand, may hold greater promise in these regards. They differ from the enumerative procedures in two respects. First, the grammars are not acquired and discarded whole, but are built up rule by rule as learning proceeds. Second, the input sentences

[1] Strictly speaking, they are not "algorithms" in the usual sense of effective procedures, since they do not compute a solution and then halt, but compute an infinite series of guesses.

do not just contribute to the binary decision of whether or not a grammar is consistent with the sample, but some property possessed by sample sentences is used as a hint, guiding the process of rule construction. Thus heuristic language learning procedures are prima facie candidates for theories of human language acquisition. They acquire language piecemeal, as children do (Brown, 1973), and they have the potential for doing so in a reasonable amount of time, drawing their power from the exploitation of detailed properties of the sample sentences instead of the exhaustive enumeration of a class of grammars.

Many heuristic procedures for acquiring rules of finite state and context-free grammars have been proposed (for examples see Biermann and Feldman, 1972; Fu and Booth, 1975; and Knobe and Knobe, 1977). The following example should give the reader the flavor of these procedures. Solomonoff (1964) suggested a heuristic for inferring recursive context-free rules from a sample, in this case with the aid of an informant to provide negative information. *Recursive rules* (not to be confused with the "recursive grammars" discussed earlier) rewrite a symbol as a string containing the original symbol, i.e., rules of the form A → BAC. They are important because they can be successively applied an infinite number of times, giving the grammar the power to generate an infinite number of sentences. An English example might rewrite the symbol for an adjective "A" as the sequence "very A". Solomonoff's learner would delete flanking substrings from an acceptable sample string, and ascertain whether the remaining string was grammatical. If so, he would sandwich that string repetitively with the substrings that were initially deleted, testing each multi-layered string for grammaticality. If they were all grammatical, a recursive rule would be constructed. For example, given the string XYZ in the original sample, the learner would test Y, then if successful, XXYZZ, XXXYZZZ, and so on. If a number of these were acceptable, the rules A → XAZ and A → Y would be coined.

Caveats concerning heuristic methods

Several points must be made about heuristic methods, lest it appear that in trading enumerative procedures for heuristic ones one gets something for nothing. First, as I have mentioned, no procedure can do better than Gold's, either in overall success or in speed, when the set of target languages consists of one of the classes in the Chomsky hierarchy. If the heuristic procedures succeed in learning some languages in a reasonable amount of time, they must take large amounts of time or fail altogether for many other ones. Thus we must again abandon the notion of a general rule learner who is constrained only by the sorts of processing or memory limits that implicitly define classes of computational procedures. Second, heuristic procedures commit the learner to assumptions not only about the target languages, but about the sentences that find their way into the sample. That is, the procedures could be fooled by using unusual or unrepresentative sets of sentences as the basis for rule construction. Consider Solomonoff's heuristic. If the target language permitted no more than three levels of embedding, the learner would have erred by constructing a rule that permitted an infinite number of embeddings. On the other hand, if the sample was a text lacking the multiply-embedded sentences that in Solomonoff's case were provided by the informant, the learner would have erred by constructing the overly-narrow rule which simply generates the original string XYZ. In the natural language case, of course, these problems are less worrisome. Not only will the child do well by "assuming" that the target language is a member of a relatively constrained set (viz., the natural languages), but he will do well in "assuming" that his sample will be a well-defined subset of the target language, not some capricious collection of sentences. Whatever its exact function may turn out to be, the dialect of speech addressed to children learning language has been found to have indisputably consistent properties across different cultures and learning environments (see Snow and Ferguson, 1977).

However, one difference between algorithmic and heuristic procedures advises caution. Whereas enumeration procedures guarantee success in learning an entire language, each heuristic at best gives hope for success in acquiring some piece of the grammar. But one can never be sure that a large collection of heuristics will be sufficient to acquire all or even a significant portion of the language. Nor can one know whether a heuristic that works well for simple constructions or small samples (e.g., the research on the construction of context-free and finite state rules cited earlier) will continue to be successful when applied to more complex, and hence more realistic tasks. In other words, in striving to meet the Developmental, Cognitive, or Time Conditions, we may be sacrificing our original goal, Learnability. The research to be discussed in the remainder of this section illustrates this tradeoff.

The computer simulation of heuristic language acquisition

Since one cannot prove whether or not a set of heuristics will succeed in learning a language, several investigators have implemented heuristic strategies as computer programs in order to observe how effective the heuristics turn out to be when they are set to the task of acquiring rules from some sample. Constructing a learning model in the form of a computer program also gives the designer the freedom to tailor various aspects of the program to certain characteristics of human language learners, known or hypothesized. Thus the theorist can try to meet several of our conditions, and is in a better position to submit the model as a *theory* of human language acquisition.

Kelley's Program

Kalon Kelley (1967) wrote the first computer simulation of language acquisition. His priority was to meet the Developmental criterion, so his program was designed to mimic the very early stages of the child's linguistic development.

Kelley's program uses a heuristic that we may call *word-class position learning.* It assumes that the words of a language fall into classes, and that each class can be associated with an absolute or relative ordinal position in the sentence. At the time that Kelley wrote the program, an influential theory ("pivot grammar", Braine, 1963) asserted that early child language could be characterized in this way. As an example of how the heuristic works, consider the following sentences:

1. (a) He smokes grass.
   (b) He mows grass.
   (c) She smokes grass.
   (d) She smokes tobacco.

A learner using the word-class position heuristic would infer that "he" and "she" belong to one word class, because they both occur as the first word of the sentence (or perhaps because they both precede the word "smokes"); similarly, "smokes" and "mows" can be placed in another word class, and "grass" and "tobacco" can be placed into a third. The learner can also infer that a sentence can be composed of a word from the first class, followed by a word from the second class, followed by a word from the third class. A learner who uses this heuristic can now produce or recognize eight sentences after having heard only four.

Kelley's program is equipped with three sets of hypotheses, corresponding to the periods in which the child uses one-, two-, and three-word utterances, respectively. The program advances from one stage to the next at arbitrary moments designated by the programmer. Its first strategy is to count the number of occurrences of various "content" words in the sample sentences; these words are explicitly tagged as content words by the "adult". It retains the most frequent ones, and can produce them as one-word sentences. In its second stage, it looks for two word classes, called "things" and "actions". Kelley assumes that children can tell whether a word refers to a thing or an action by the non-linguistic context in which it was uttered. To model this assumption, his program guesses arbitrarily that a particular word is in one or the other class, and has access to its "correct" classification. If the guess is correct, it is strengthened as a hypothesis; if incorrect, it is weakened. At the same time, the program tabulates the frequency with which the word classes precede or follow each other, thereby hypothesizing rules that generate the frequent sequences of word classes (e.g., S → thing action; S → thing thing). Like the hypotheses that assign words to classes, these rules increase or decrease in strength according to how frequently they are consistent with the input sentences. In its third state, the program retains its two word classes, and adds a class consisting of two-item sequences (e.g., thing-action) from the previous stage. As before, it accumulates evidence regarding which of these classes can occur in which sentence positions relative to one another, thereby hypothesizing rules that generate frequent sequences of classes (e.g., S → thing-action thing). A separate feature of the program is its ability to learn the "functions" of the individual sentence constituents, such as which is the subject and which is the predicate. As before, the program learns these by making rather arbitrary guesses and checking them against the "correct" answer, to which it has access.

An evaluation

Though Kelley's program was a brave first attempt, it is unsatisfactory on many counts. For one thing, children seem unaffected by the frequency of syntactic forms in adult speech (Brown, 1973), whereas frequency of input forms is the very life-blood of Kelley's learning procedure. Second, the role of the "correct" structural descriptions of sentences given to the program is puzzling. Kelley intends them to be analogous to the child's perception that a word uttered in the context of some action is an "action" word, that a part of a sentence denoting an object being attended to is the "subject" of the sentence, and so on. But in the context of the program, this is reduced to the trivial process of guessing the class or function of a word, and being told whether or not the guess is correct. I will review more systematic attempts to simulate perceptual and pragmatic clues in Sections VI-VIII. Finally, the heuristics that the program uses are inadequate to advance beyond the three-word stage since, as we shall see, natural languages cannot be characterized by sequences of word classes. In any case, one must question whether there is really any point in doing simulations that address themselves only to the Developmental Condition. The early stages of language development can easily be accounted for by all sorts of ad hoc models; it is the acquisition of the full adult grammar that is the mystery.

The Distributional Analysis Heuristic

The problem with the word-class position heuristic when it is applied to learning natural languages is that it analyzes sentences at too microscopic a level. It is practically impossible to state natural language regularities in terms of contiguous word classes in sentences. Consider the following sentences:

2. (a) That dog bothers me.
(b) What she wears bothers me.
(c) Cheese that is smelly bothers me.
(d) Singing loudly bothers me.
(e) The religion she belongs to bothers me.

In the different sentences, the word "bothers" is preceded by a noun, a verb, an adjective, an adverb, and a preposition. Clearly there is a generalization here that an astute learner should make: in all the sentences, "bothers" is preceded by a noun phrase. But noting that certain word classes precede "bothers" will not capture that generalization, and will only lead to errors (e.g., "Loudly bothers me").

A more general heuristic should look for more flexible contexts than either ordinal position in a sentence or position relative to an adjacent item, and should define classes more broadly, so that each class can consist of strings of words or subclasses instead of single words. Kelley's program moved in this direction in its third stage. Heuristics of this sort are often called *distributional analysis* procedures (see Harris, 1964), and exploit the fact that in context-free languages, the different instantiations of a grammatical class are interchangeable in the same linguistic context. Thus it is often a good bet that the different strings of words that all precede (or follow, or are embedded in) the same string of words all fall into the same class, and that if one member of such a class is found in another context, the other members of that class can be inserted there, too. Thus in sentences 2(a-e), a distributional analysis learner would recognize that all strings that preceed "bothers me" fall into a class, and that a member of that class followed by the phrase "bothers me" constitutes a sentence. If the learner then encounters the sentence "That dog scares me", he can place "scares me" and "bothers me" into a class, and "scares" and "bothers" into a subclass. If he were to encounter "Sol hates that dog", he could place all the noun phrases in the first class after the phrase "Sol hates". By this process, the learner could build up categories at different levels of abstraction, and catalogue the different ways of combining them in sentences.

Problems with distributional analysis

There are several hurdles in the way of using distributional analysis to learn a natural language. First, it requires a great many sets of minimally-contrasting sentences as input. We know that American children often do hear closely-spaced sets of sentences with common constituents (e.g., Brown, Cazden, and Bellugi, 1969; Snow, 1972; see Snow and Ferguson, 1977), but we do not know whether this pattern is universal, nor whether it occurs with enough grammatical constituents to determine uniquely every rule that the child can master. Second, a distributional analysis of a sample of a natural language is fraught with the possibility for serious error, because many words belong to more than one word class, and because virtually any subsequence of words in a sentence could have been generated by many different rules. For example, sentences 3(a-d)

3. (a) Hottentots must survive.
   (b) Hottentots must fish.
   (c) Hottentots eat fish.
   (d) Hottentots eat rabbits.

would seduce a distributional analysis learner into combining heterogeneous words such as "must" and "eat" into a single class, leading to the production of "Hottentots must rabbits", "Hottentots eat survive", and other monstrosities.

Finally, there is a combinatorial explosion of possibilities for defining the context for a given item. Given *n* words in a sentence other than the item of interest, there are $2^n - 1$ different ways of defining the "context" for that item – it could be the word on the immediate right, the two words on the immediate left, the two flanking words, and so on. In combination with the multiple possibilities for focusing on an item to be generalized, and with the multiple ways of comparing items and contexts across large sets of sentences, these tasks could swamp the learner. However by restricting the types of contexts that a learner may consider, one can trade off the first and third problems against the second. An extremely conservative learner would combine two words in different sentences into the same class only if all the remaining words in the two sentences were identical. This would eliminate the explosion of hypotheses, and sharply reduce the chances of making overgeneralization errors, but would require a highly overlapping sample of sentences to prevent undergeneralization errors (for example, considering every sentence to have been generated by a separate rule). Siklóssy (1971, 1972) developed a model that relies on this strategy. On the other hand, a bolder learner could exploit more tenuous similarities between sentences, making fewer demands on the sample but risking more blunders, and possibly having to test for more similarities. It is difficult to see whether there is an "ideal" point along this continuum. In any case no one has reported a successful formalization or computer implementation of a "pure" distributional analysis learner. Instead, researchers have been forced to bolster a distributional analysis learner with various back-up techniques.

An "Automated Linguist"

Klein and Kuppin (1970) have devised what they call "an automatic linguistic fieldworker intended to duplicate the functions of a human fieldworker in learning a grammar through interaction with a live human informant". Though never intended as a model of a child, "Autoling", as they call it, was the most ambitious implementation of a heuristic language learner, and served as a prototype for later efforts at modelling the child's language learning (e.g., Anderson, 1974; Klein, 1976).

Use of distributional analysis

The program is at heart a distributional analysis learner. As it reads in a sentence, it tries to parse it using the grammar it has developed up until that point. At first each rule simply generates a single sentence, but as new sentences begin to overlap with old ones, the distributional heuristics begin to combine words and word strings into classes, and define rules that generate sequences of classes and words. Out of the many ways of detecting similar contexts across sentences, Autoling relies most heavily on two: identical strings of words to the left of different items, and alternating matching and mismatching items.

Generalizing rules

Autoling also has heuristics for generalizing rules once they have been coined. For example, if one rule generates a string containing a substring that is already generated by a second rule (e.g., X → ABCD and Y → BC), the first rule is restated so as to mention the left-hand symbol of the second rule instead of the substring (i.e., X → AYD; note that this is a version of Solomonoff's heuristic). Or, if a rule generates a string composed of identical substrings (e.g., X → ABCABC), it will be converted to a recursive pair of rules (i.e., X → ABC; X → XABC). Each such generalization increases the range of sentences accepted by the grammar.

Taming generalizations

In constructing rules in these ways, Autoling is generalizing beyond the data willy-nilly, and if left unchecked, would soon accept or generate vast numbers of bad strings. Autoling has three mechanisms to circumvent this tendency. First, whenever it coins a rule, it uses it to generate a test string, and asks the informant whether or not that string is grammatical. If not, the rule is discarded and Autoling tries again, deploying its heuristics in a slightly different way. If this fails repeatedly, Autoling tries its second option: creating a transformational rule. It asks its informant now for a correct version of the malformed string, and then aligns the two strings, trying to analyze the correct string into constituents similar to those of the malformed string. It then generates a rule that transforms the malformed into the correct string, permuting or deleting the most inclusive common constituents. As before, it uses the new transformation to generate a test string, and asks the informant for a verdict on its grammaticality, discarding the rule and trying again if the verdict is negative. Finally, if nothing succeeds, the entire grammar self-destructs, and the heuristics begin again from scratch on the entire collection of acceptable sentences, which have been retained since the beginning of the learning session.

An evaluation

Autoling was not meant to be a model of the child, and needless to say, it is far from one. Unlike children, it scans back and forth over sentences, makes extensive use of negative feedback and corrections from an informant (cf., Brown *et al.*, 1969), tests each new rule methodically, remembers every sentence it hears, and gives up and restarts from scratch when in serious trouble. But it is important as a vivid illustration of the pitfalls of building a language learning model around a collection of heuristics. It is bad enough that Autoling resembles one of Rube Goldberg's creations, with its battery of heuristics (only a few of which I have mentioned), its periodic checkings and recheckings for overlapping, redundant, or idle rules, its various cleanup routines, its counters tabulating its various unsuccessful attempts, and so on. But even with all these mechanisms, Autoling's success as a language learner is very much in doubt. Klein and Kuppin do present records of the program successfully inducing grammars for artificial languages such as a set of well-formed arithmetic expressions. But as an illustration of its ability to learn a natural language, they present a rather unparsimonious grammar, constructed on its second attempt, which generates a finite fragment of English together with a variety of gibberish such as "need she" and "the want take he". Klein and Kuppin are simply unable to specify in any way what Autoling can or cannot learn. Thus Autoling – and, I would argue, any other attempt to model grammar acquisition via a large set of ad hoc heuristics – does not seem a promising start for an adequate theory of language learning. Not only does it violate the Developmental, Cognitive, and Input Conditions, but it does not even come close to meeting the Learnability Condition – the chief motivation for designing learning simulations in the first place.

VI. Semantics and Language Learning

I have postponed discussing the role of semantics in language learning for as long as possible, so as to push the purely syntactic models as far as they can go. But the implausibility of both the enumerative and the heuristic learners seems to indicate that the time has come.

The "Cognitive Theory" of Language Learning

The semantic approach to language learning is based on two premises. First, when children learn a language, they do not just learn a set of admissible sentences; they also learn how to express meanings in sentences. Second, children do not hear sentences in isolation; they hear them in contexts in which they can often make out the intended meanings of sentences by non-linguistic means. That is, they can see what objects and actions are being referred to in the sentences they hear, and they can discern what their parents are trying to communicate as they speak. (Kelley incorporated a version of this assumption into his model.) An extremely influential theory in developmental psycholinguistics (often called the "Cognitive Theory") asserts that children learn syntax by inferring the meanings of sentences from their non-linguistic contexts, then finding rules to convert the meanings into sentences and vice-versa (Macnamara, 1972; Schlesinger, 1971). Several considerations favor the Cognitive Theory. The first (though rarely cited) consideration is that semantic information can substitute for information about non-sentences to make classes of languages formally learnable. The second is that there is some empirical evidence that both children and adults use semantic information when they learn syntactic rules. The third consideration is that this task is thought to be "easier" than inferring a grammar from a set of strings alone, because the mental representations corresponding to sentence meanings are thought to resemble the syntactic structures of sentences. I will discuss each justification for the semantic approach in turn.

Learnability with Semantic Information

John Anderson (1974, 1975, 1976) has described a semantic version of Gold's language acquisition scenario, formalizing an earlier speculation by Clark (1973). First, he assumes that whatever "sentence meanings" are, they can be expressed in a formal symbolic notation, and thus can be put into one-to-one correspondence with the set of natural numbers by the mathematical technique known as "Gödelization". Second, he assumes that a natural language is a function that maps sentences onto their meanings, or equivalently, well-formed strings onto natural numbers, and vice-versa. (In contrast, we have been assuming that natural languages are functions that map strings onto the judgments "grammatical" and "non-grammatical", or equivalently, "1" and "0".) Third, he assumes that children have access to a series of pairs consisting of a sentence and its meaning, inferred from the non-linguistic context.

The child's task is to identify in the limit a function which maps sentences onto their meanings.

Recall that Gold (1967) proved that the class of primitive recursive functions, which map strings onto numbers, is learnable provided that the learner has eventual access to all number-string pairs. For Gold, the numbers represented the trial number or time since the start of learning, but in Anderson's model, the numbers correspond to sentence meanings. The learner enumerates the primitive recursive functions, testing each one against the sample of sentence-meaning pairs, retaining a function if it is consistent with the sample (see Figure 1). In this way the learner will identify the function (and hence the language) in the limit, since all incorrect functions will be rejected when they pair a meaning with a different string than the one in the sample.

Although in this version the learner can be proved to succeed without requiring information as to what is not a sentence, all of Gold's other conclusions remain in force. It will take the learner an astronomical amount of time until he arrives at the correct function, but there is no quicker or more successful method, on the whole, than enumerating functions one by one. By suitably restricting the learner's hypothesis space, learning time can be reduced, and by using heuristic procedures that exploit properties of individual meaning-sentence pairs, it can be reduced even further. But once again the learner ceases to be a multipurpose rule learner – he makes tacit assumptions about the syntax of the target language, about the way that meanings are mapped onto strings, and about the representativeness of the meaning-sentence pairs in the sample at a given time. He will fail to learn any language that violates these assumptions. As Chomsky (1965) has noted, the hypothesis that the child uses semantics in learning syntax is in some senses stronger, not weaker, than the hypothesis that sentences alone are used.

Evidence for the Cognitive Theory

Cognitive development and language acquisition

Two sorts of evidence have been martialled in support of the view that humans base their learning of syntax upon their conceptualization or perception of the meanings of sentences. The first consists of various correlations between language development and cognitive development, which are thought to imply that the non-linguistic mental representations available to the child constrain the linguistic hypotheses that he will entertain. For example, the early two- and three-word utterances of children seem to reflect closely certain semantic relations such as agent-action, possessor-possessed, etc. (Bowerman, 1973; Brown, 1973; Schlesinger, 1971). As well, the "cognitive complexity" of the semantic functions underlying various grammatical rules has been shown to predict in a rough way the order of the child's mastery of those rules (Brown, 1973). Similarly, it has been found that some syntactically simple rules (such as the conditional in Russian) are not acquired until the underlying semantic functions (in this case, implication) have been mastered (Slobin, 1973).

Semantics and artificial language learning

The second sort of evidence comes from a set of experiments in which adult subjects are required to learn artificial languages, that is, they must learn to discriminate grammatical from ungrammatical test strings as defined by a grammar concocted by the experimenter. In early experiments of this type (e.g., Miller, 1967), where subjects saw various strings of nonsense syllables, even the simplest grammars were extremely difficult for the subjects to learn. However, in a famous set of experiments, Moeser and Bregman (1972, 1973) presented some subjects with a sample of strings, and other subjects with a sample in which each string was paired with a picture of geometric forms such that the shapes, colors, and spatial relations of the forms corresponded to the words and syntactic relations in the sentences (that is, the pictures were intended to serve as the semantic referents of the strings). After more than 3000 strings had been presented, the subjects who saw only strings failed utterly to discriminate grammatical from ungrammatical test strings, while those who saw strings and pictures had no trouble making the discrimination. This finding has led many theorists to conclude that it is intrinsically easier for humans to learn syntactic rules if they use semantic information in addition to sentences.

However Anderson (1974, 1975) has pointed out that semantics-based learners, including the subjects in Moeser and Bregman's studies, learn by virtue of specific assumptions they make about the way the target language uses syntactic structures to express semantic relations. For example, he notes that natural languages require an adjective to predicate something about the referent of the noun in its own noun phrase, never a noun in another noun phrase in the sentence. That is, in no natural language could a phrase such as "the blue stripes and the red rectangle" refer to an American flag, even though the sentences of such a language might be identical to the sentences of (say) English, and the semantic relations expressible in that language might be identical to those expressible in (say) English. Anderson performed an experiment in which subjects saw strings of English words (referring to shapes, colors, and spatial relations) generated by an artificial grammar. A second group saw the same strings paired with pictures in such a way that each adjective in the sentence modified the noun in its phrase; a third group saw the same strings and pictures, but they were paired in such a way that each adjective modified a noun in a different phrase (like our example with the flag).

Only the second group of subjects, with the "natural semantics", were later able to discriminate grammatical from ungrammatical test strings. Thus, Anderson argues, it is not the availability of semantic information per se that facilitates syntax learning in humans, but semantic information that corresponds to the syntactic structures in the target language in some assumed way.[2] These correspondences will be explained in the next section, in which semantics-based learning heuristics are discussed.

Heuristics that use Semantics

The most important fact about the natural language acquisition task is that the units composing linguistic rules are abstract, and cannot be derived from sample strings in any simple way. The problem with distributional analysis was that these units or "constituents" do not uniquely reveal themselves in the patterns of sentence overlappings in a sample. However, if the semantic representation of a sentence corresponds in a fairly direct way to the syntactic description of that sentence, semantic information can serve the same purpose as distributional regularities. The syntactic structure of a sentence in a context-free or context-sensitive language can be depicted as a tree, with each node representing a constituent, and the set of branches emanating from a node representing the application of a rule rewriting that constituent as a sequence of lower-order constituents. Similarly, the mental representational structures corresponding to percepts and sentence meanings are also often represented as trees or similar graph structures (e.g., Anderson and Bower, 1973; Norman and Rumelhart, 1975; Winston, 1975). The top nodes of such trees usually correspond to logical propositions, and the branches of these trees correspond to the breakdown of propositions into their subjects and predicates, and to the successive breakdown of the subject and predicate into concepts and relations, or into further propositions. If the tree representing a sentence meaning is partially isomorphic to the constituent structure of the sentence, presumably there is a way that a child can use the meaning structure, which by assumption he has, to discern the constituent structure of the sentence, which he does not have. Anderson (1974, 1975, 1977) has demonstrated precisely how such heuristics could work. In the following paragraphs I shall explain the operation of these heuristics; then, in Section VII, I shall show how Anderson has embodied these heuristics in a computer model of the language learner.

[2] Of course, in this particular case the assumption about semantics and syntax need not have been innate, since the subjects' tacit knowledge of English could have constrained their hypotheses.

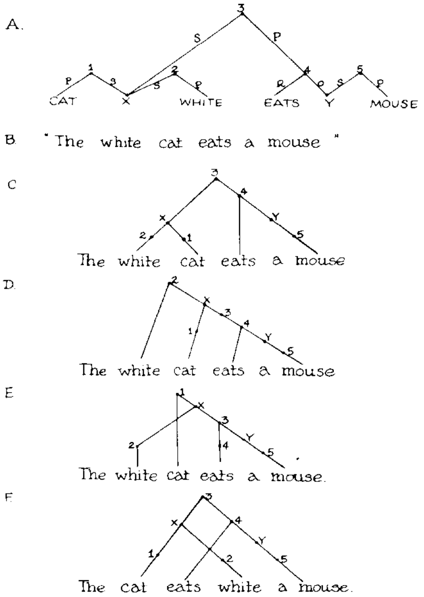
Using semantics to delineate constituents: the Tree-fitting heuristic

This heuristic begins with the assumption that the child knows the meaning of all the "content" words in the sentence, that is, he knows to which concept node in the meaning structure each word corresponds. The learner matches the concepts in the meaning structure to the words in the sentence, and attempts to fit the tree structure for the meaning onto the sentence, spatially rearranging the nodes and branches as necessary but preserving all links between nodes. The learner now has a tree-structure for the sentence, and can deduce what the constituents are and how the rules of the grammar rewrite the major constituents as sequences of minor ones.

An example will make this heuristic clearer. Say the child saw a white cat eating a mouse. His perceptual system might construct the propositions "X is a CAT", "X is WHITE", "Y is a MOUSE", and "X EATS Y", which can be depicted as a single tree-structure like the one in Figure 3(a). Say the child simultaneously heard the string of words "the white cat eats a mouse". By matching the word "white" onto the concept "WHITE" (and so on for the other words), reversing the order of the respective links to "CAT" and to "MOUSE", and straightening out continuous series of links, the child can arrive at the tree-structure for the sentence which is depicted in Figure 3(c). He can then hypothesize rules specifying that a sentence can be broken down into two constituents, that one constituent can be broken down into a class containing the word "white" and another containing the word "cat", and that the second main constituent can be broken down into the word "eats" and a constituent containing a class containing the word "mouse". Furthermore, the child can construct rules translating syntactic constituents into semantic propositions and vice-versa. In this example, he could hypothesize that the first major constituent of a sentence refers to some individual that is the subject of an underlying proposition, the first word class in this constituent refers to some property predicated of that individual, and so on.

The problem with this heuristic is that there are usually many ways to fit a semantic structure onto a string of words, only one of which will correspond to the correct breakdown of the sentence into its syntactic constituents. For example, nothing would have prevented the child in our example from constructing the syntactic trees depicted in Figures 3(d) and (e) instead of the one in Figure 3(c). Anderson has proposed two mechanisms by which the heuristic could "know" the best way to fit the semantic tree onto the string. First, the learner must know which node of the semantic tree should be highest in the syntactic tree, in order to distinguish between the possibilities represented in Figures 3(c) and (d). This corresponds to knowing the main proposition of the sentence, that is, what is the major topic of the sentence and what is the major thing being asserted of it. Anderson suggests that

Figure 3. *Semantic structure (a) to be fitted onto the string (b) in various ways by the Tree-fitting heuristic. In this formalism for semantic structures (HAM; Anderson and Bower, 1973), S = subject, P = predicate, R = relation, O = object, X and Y represent individuals, and capitalized terms are concepts, which correspond to words.*


A.
3
S
P
1
P
S
2
S
P
4
R
O
5
S
P
CAT
X
WHITE
EATS
Y
MOUSE
B
"The white cat eats a mouse"
C
3
4
X
Y
2
1
5
The white cat eats a mouse
D.
2
X
3
4
1
Y
5
The white cat eats a mouse
E
1
X
3
Y
2
4
5
The white cat eats a mouse.
F
3
X
4
Y
1
2
5
The cat eats white a mouse.

this pragmatic information is communicated to the child during his normal interactions with adults; in other words, the social and communicative context in which a sentence is uttered makes it clear what the adult intends to assert about what (see Bruner, 1975, for supporting arguments and evidence). For the tree-fitting heuristic, this means that one of the propositions in the semantic structure is tagged as the "principal" one, and its node will be highest when the semantic tree is fitted onto the string of words. The nodes connected to this "root" node by one link are placed one level lower, followed by the nodes connected to the root by two links, and so on. Thus if the main proposition concerns what the cat did to the mouse, the heuristic will fit the tree depicted in Figure 3(c) onto the string. On the other hand, if it is the whiteness of the mouse-eating cat that is being asserted (e.g., "white is the cat that eats the mouse"), the heuristic will fit the tree depicted in Figure 3(d) onto the string.

The second constraint on the heuristic is that no branches be allowed to cross. Thus the heuristic would be prohibited from fitting the tree depicted in Figure 3(e) onto the string. No set of context-free rules can generate a tree like this, and in fact what the constraint does is prevent the heuristic from constructing trees from which no context-free rules can possibly be derived. Thus this constraint, which Anderson calls the *Graph Deformation Condition*, will prevent the learner from learning languages that use certain rules to transform meaning structures into sentences. For example, it cannot learn a language that could express the semantic structure in Figure 3(a) by the string of words "the cat eats white a mouse". Nor could it learn the "unnatural semantics" language that the subjects in Anderson's experiment failed to learn. In each case it would be unable to fit the semantic structure onto the string without crossing branches, as Figure 3(f) shows. In general, the heuristic is incapable of learning languages that permit elements from one constituent to interrupt the sequence of elements in another constituent. As Anderson argues, this is a particularly telling example of how a semantics-based heuristic in effect assumes that the language it faces maps meanings onto sentences only in certain ways. In this case, the Tree-fitting heuristic "assumes" that the language meets the Graph Deformation Condition. Anderson believes that natural languages obey this constraint for the most part, and that both children and adults (such as his experimental subjects) tacitly assume so as they use the Tree-fitting heuristic. I will discuss these claims in Section VII.

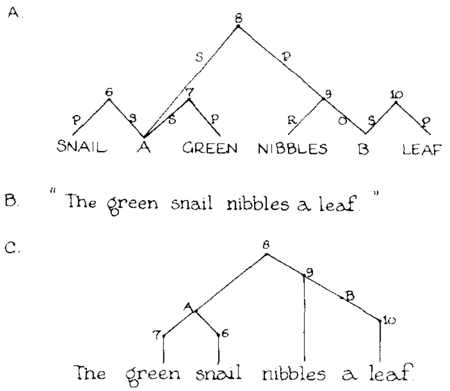
Using semantics to generalize rules

Once the learner has broken down sentences into their constituents and hypothesized the corresponding rewrite rules, he must combine rules that have been derived from different sentences – otherwise he is left with one set of rules for each sentence, not much better than a learner who simply memorized the sentences whole. Rule-merging is a particularly rocky step for distributional analysis heuristics(as sentences 3(a-d) showed), since sentences from natural languages provide countless temptations to merge dissimilar constituents owing to the syntactic ambiguity of most short substrings. Klein and Kuppin's program tentatively merged rules with overlapping constituents, used the newly-merged rules to generate a sentence, and submitted the sentence to the informant for approval before it would declare the merger permanent. But this is an unrealistic way to keep overgeneralizations in check. Not only do children not have access to such an informant, but even if they did, it is unlikely that the Autoling strategy would work as required. A merged rule can usually generate many sentences, sometimes an infinite number, so the knowledge that one string is acceptable does not mean that all the strings generated by the rule will be acceptable.

However information in the semantic representation might be used instead to decide whether rules can safely be merged. First, Anderson suggests that words in the same positions in different sentences whose concepts have identical roles in the semantic structure can be merged into one class. For example, say the learner, after processing the meaning–sentence pair in Figure 3(c), encountered the sentence "The green snail nibbles the leaf", together with its semantic structure, as shown in Figure 4(a) and (b). After fitting the semantic tree onto the string (see Figure 4(c)) and deriving the corresponding rules, the learner can use the similarities between the semantic representations in Figures 3(a) and 4(a) to merge the two sets of rules. For example, "EATS" in Figure 3(a) corresponds to the "relation" branch of the predicate of the main proposition, and so does "NIBBLES" in Figure 4(a). The learner can then merge the corresponding words into one class, and by similar means can merge "white" and "green", "eat" and "snail", and so on.

Now the learner must recognize that the higher constituents in the two sentences can also be merged, such as the ones embracing "the white cat" and "the green snail". Anderson suggests a double criterion for when to merge higher-order constituents: they must decompose into identical sub-constituents, and they must serve the same semantic role. In this example, both are satisfied: the word classes in the two constituents have already been merged, and both constituents serve as the subject of their respective main propositions. Once all the parallel constituents in the two sentences have been merged, the learner will end up with a grammar that generates sixteen different sentences: "the green cat eats a leaf", "the white snail nibbles a mouse",

Figure 4. *Semantic structure* (A), *string* (B), *and tree* (C) *which, in conjunction with Figure 3, illustrate the Semantics-Induced Equivalence Heuristic.*


A.
8
S
P
6
7
9
10
P
S
S
P
R
O
S
P
SNAIL
A
GREEN
NIBBLES
B
LEAF
B.
" The green snail nibbles a leaf. "
C.
8
9
A
B
7
6
10
The green snail nibbles a leaf.

and so on. Anderson calls this heuristic (and the putative property of natural languages that it exploits) *Semantics-Induced Equivalence of Syntax.* He asserts that the heuristic exploits the tendency of natural languages always to use the same syntactic construction to express a particular semantic relation within a given higher-order constituent. Whether or not this claim is true of English will be discussed in Section VII.

It is interesting to note that the Semantics-Induced Equivalence of Syntax heuristic is neither more nor less conservative, on the whole, than Distributional Analysis. Each will try to merge in situations where the other would not. Thus the Distributional Analysis heuristic would make no generalization embracing the sentences in Figures 3 and 4, since they share no content words. Instead it would have to wait until some sentence like "the green snail eats a mouse" appeared. On the other hand, the Semantics-Induced Equivalence heuristic, upon encountering the sentence "the white cat eats slowly", would not merge "slowly" with "the mouse" (as would Distributional Analysis), since "MOUSE" and "SLOWLY" would not have the same roles in their semantic structures. It should be clear from these examples that the Semantics-Induced Equivalence heuristic will, in general, make the wiser generalization.

VII. Anderson's Language Acquisition System

The Computer Simulation of Semantics-based Heuristic Language Acquisition

Heuristics that exploit syntax-semantics correlations, like those that exploit properties of sentences alone, are often implemented as computer programs (Anderson, 1974, 1975, 1977; Fabens and Smith, 1975; Klein, 1976; Klein and Rozencvejg, 1974; McMaster, Sampson, and King, 1976; Reeker, 1976; Siklóssy, 1971, 1972). In a sense, these programs are incarnations of the informal Cognitive Theories of the Schlesinger and Macnamara sort. As such, they serve as a testing ground for the adequacy of those theories, especially at meeting the Learnability Condition, and can also contribute to the goal of specifying more precisely and explicitly the mechanisms that these theories implicate. Unfortunately, many of the programs that have been developed succumb to the same syndrome that afflicted Klein and Kuppin's model: unreasonable assumptions about the learner and the information available to him, ad hoc and unparsimonious learning mechanisms, and dubious success at learning. For example, the program of Fabens and Smith (1975) modifies its rules in accordance with environmental approval and disapproval, which Brown and Hanlon (1970) have shown is probably irrelevant to the learning of syntax. Other programs (e.g., Klein, 1976; Reeker, 1976; Siklóssy, 1971,

1972) avoid this device but only learn to produce meager, ill-defined fragments of natural languages, often generating many non-sentences at the same time. The exception among these efforts is Anderson's Language Acquisition System (LAS; 1974, 1975, 1977). As we have seen, Anderson has carefully defined certain heuristics that his program employs and the properties of natural languages that make these heuristics useful. As well, the program can acquire well-defined infinite subsets of natural languages, its semantic representations have an independent theoretical motivation, and it avoids for the most part psychologically unrealistic strategies. For these reasons, I will discuss only Anderson's simulation from among the many that have been reported (which in any case rely on heuristics remarkably similar to the ones Anderson uses).

How LAS works

General architecture

LAS uses a formalism for semantic representations that Anderson has used elsewhere as a theory of information representation in long term memory (the Human Associative Memory system (HAM) of Anderson and Bower, 1973). Its grammar is in the form of an Augmented Transition Network (ATN), which is held by many to be a plausible model of human language processing (see Kaplan, 1975). The ATN that LAS uses corresponds rule-for-rule to a context-free grammar, but can be incorporated more easily into a left-to-right sentence recognizer or producer. LAS has a subroutine corresponding to sentence production, which uses the ATN to convert a semantic structure into a sentence. It also has a subroutine that corresponds to sentence comprehension, which uses the ATN to convert a sentence into its semantic structure. Finally, it has a learning program that uses pairs consisting of semantic structures and sentences to build the ATN piece-by-piece. The latter program is the one of interest here.

Like Kelley's and Klein and Kuppin's programs, LAS is driven by the comprehension process. It tries to interpret a sentence from left-to-right with its current grammar, and alters parts of the grammar if it fails. If a particular rule gets the learner part way in interpreting a sentence, it is the one that will be expanded. LAS also forgets the exact sentences that it hears, so that a sentence contributes to grammatical development only in the way that it alters the grammar as it is being understood. These features give LAS a psychologically realistic flavor compared to other models I have discussed.

Use of the Tree-fitting heuristic

When LAS receives its first sentence-meaning pair, there is no grammar to interpret it yet, so it must build the first pieces of the grammar relying entirely on the Tree-fitting heuristic. But in general, the HAM structure representing the learner's perception of the situation in which the sentence has been uttered is not really suitable for fitting onto the string right away. It contains too many sentence-irrelevant propositions, and has no way of indicating the proposition corresponding to the principle assertion of the sentence (see Section VI). Thus the program is forced to compute an intermediate representation, called the *Prototype Structure,* which omits propositions whose concepts have no counterparts among the words of the sentence, and highlights the principle proposition (in line with supposed pragmatic cues). It is this Prototype structure, not the meaning structure itself, that the Tree-fitting heuristic tries to fit onto the string of words. Once an acceptable tree has been derived by the heuristic, LAS constructs ATN *arcs,* each one corresponding to a left-to-right sequence of constituents composing a higher constituent, and the corresponding rules that map these syntactic constituents onto their semantic counterparts.

Use of the semantics-based equivalence heuristic

When subsequent pairs come in, LAS tries to interpret the strings using all its rules simultaneously. Using the Semantics-Induced Equivalence heuristic, it unites into a single class words whose concepts serve the same role in their respective HAM structures. Similarly, it merges any two arcs (i.e., higher-order constituents) that simultaneously assign the same semantic role to their respective sentence constituents. These mechanisms were discussed in Section VI. In addition, LAS will merge two arcs if one is a proper subsequence of another, as long as they both specify the same semantic role. For example, assume that LAS has induced an arc that will parse sequences like "the mouse" in Figure 3, and that it is forced by a subsequent sentence to construct an arc that will parse "the mouse that nibbles the house". Then the old arc will be swallowed into the new one automatically (with the last four words marked as "optional"). In this way, LAS can construct recursive rules, allowing it to generate infinite languages. In the present example, it would construct a low-level arc to parse "the house"; however, this substring can already be parsed with the higher-level arc built to parse "the mouse that nibbles the house" (since "mouse" and "house" would presumably be merged, and the last four words are marked as optional). Consequently it would merge the two arcs, ending up with the recursive arc corresponding to the rule "*noun phrase* → the *noun* that nibbles *noun phrase*". Now it can generate "the mouse that nibbles the cat that eats the mouse that nibbles the house" and so on.

Finally, LAS has a special heuristic with which it handles the so-called "grammatical morphemes" such as articles, auxiliaries, relative pronouns, and so on, which have no direct counterparts in the semantic representations. This heuristic will be discussed in a later paragraph.

Learning powers of LAS

How well does LAS do? Anderson presents several examples in which LAS is faced with artificial languages or fragments of natural languages, all context-free, which can be used to describe arrangements of two-dimensional shapes of various colors and sizes. In all cases LAS succeeded in acquiring a grammar for the language, including infinitely large subsets of English and French, after taking in 10-15 meaning-sentence pairs. For example, it could handle sentences like "the large blue square which is below the triangle is above the red circle which is small", and other sentences using these grammatical constructions. Anderson conjectures that LAS could learn any context-free language with a semantic system that respected the Graph Deformation Condition and the Semantics-Induced Equivalence of Syntax Condition.

An Evaluation of LAS

LAS is unquestionably an impressive effort. Anderson is alone in showing how a learner with semantics-based heuristics can succeed in learning chunks of natural languages in a plausible manner. Furthermore, there are possibilities for extending the powers of LAS. If LAS were built like Winograd's (1972) program to converse with another speaker instead of receiving sentences passively, it would have representational structures that conceivably could be useful in acquiring rules for interrogatives, conditionals, imperatives, and so on. And if it had a more childlike semantic representational system, which categorized the world into actors, actions, and recipients of actions, possessors and possessed, objects and locations, and so on, its linguistic abilities might even resemble those of young children (cf., Brown, 1973). By enriching the semantic system gradually, it might even be possible to generate a sequence of stages parallel to the child's linguistic development, which would be a unique accomplishment among formal models of language learning (outside of Kelley's limited attempts). Of course, all of this remains to be shown.

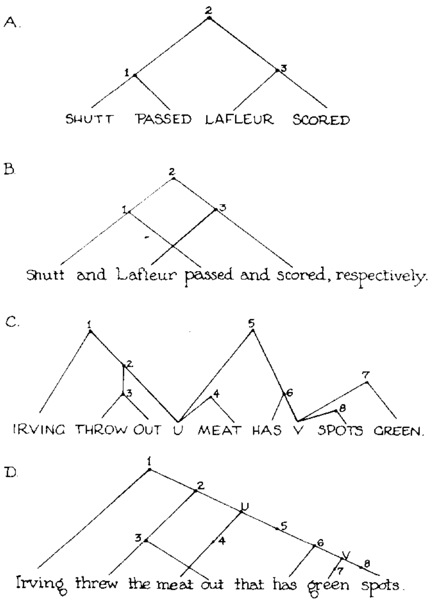
In any case, rather than spelling out the various ways that LAS can be extended, I shall focus in this section on the *limits* of LAS's abilities, on Anderson's claim that "the weakness of LAS... is sufficiently minor that I am of the opinion that LAS-like learning mechanisms, with the addition of some correcting procedures, could serve as the basis for language learning" (1977, p. 155-156). Since LAS is an incarnation of the currently popular Cognitive Theory of language learning, Anderson's claim is an important one. If true, it would support the contention that the child's perceptual and cognitive representations are sufficiently rich data structures to support language acquisition (e.g., Bowerman, 1973; Sinclair-de Zwart, 1969; Schlesinger, 1971), obviating the need for innate language-specific data structures (e.g., Chomsky, 1965; Fodor, 1966; McNeill, 1966). On this view, the innate constraints on the learner derive only from his cognitive representational structures and, as Anderson points out, his tacit assumptions about how these correspond to syntactic structures. For this reason I will examine LAS's abilities in some detail. In particular, I shall scrutinize Anderson's central claim, that most syntactic rules can be derived from distinctions made at the semantic level, while the rest can be derived with the help of a few miscellaneous heuristics.

Do natural languages obey the Graph Deformation Condition?

This condition, on which the Tree-fitting heuristic depends, decrees in effect that natural languages must be context-free, a conclusion that Anderson explicitly supports (despite its near-universal rejection by linguists). There are a number of natural language constructions which cross branches, and Anderson must find reason to dismiss them as counter-examples to the omnipotence of the Tree-fitting heuristic. One example is the "respectively" construction. As Figures 5(a) and (b) show, the semantic structures for these sentences cannot be fitted onto the strings without branches crossing. A second example can be found in languages that indicate semantic roles by case markers instead of by word order (e.g., Russian, Latin, Wolbiri). In these languages it is possible for an element that belongs to one phrase to interrupt a sequence of elements in a second phrase, provided that the intruding element is suitably marked as belonging to its phrase. Anderson cites both these counter-examples, and argues that they are atypical constructions, possibly acquired by special problem-solving strategies outside the normal language induction mechanisms. While the rarity of constructions of the "respectively" sort make this conclusion tenable for these constructions, it is less easy to forgive the paucity of mechanisms in LAS for acquiring case-inflection rules, prevalent in languages other than English, which naturally give rise to constructions with crossing branches.

A second class of counter-examples consists of discontinuous elements, which give rise to crossing syntactic dependencies in a sentence. For example, in the sentence "Irving threw the meat out that had green spots", the phrase "the meat" is part of a constituent that includes "that had green spots", whereas the word "threw" is part of a constituent that includes the word "out". Figure 5(c) and (d) show how these branches must cross (similar crossing dependencies can occur with auxiliary and tense morphemes under certain analyses, see Gross, 1972). Anderson exempts the Tree-fitting heuristic from having to deal with such constructions on the grounds that they involve "non-meaning bearing morphemes" which are outside its province. But this is not quite true – the morpheme "out" in the sentence in Figure

Figure 5. *Violations of the Graph Deformation Condition.*


A.
2
1
3
SHUTT PASSED LAFLEUR SCORED
B
2
1
3
Shutt and Lafleur passed and scored, respectively.
C.
1
5
2
3
4
6
7
8
IRVING THROW OUT U MEAT HAS V SPOTS GREEN.
D
1
2
U
5
3
4
6
V
8
7
Irving threw the meat out that has green spots.

5(d) conveys a different meaning than would the morphemes "up" or "around" or "down" if one were substituted in its place. But it is not clear how the morpheme "out" would have been mapped onto the semantic structure in the first place – if "THROW-OUT" were represented as a unitary node, and the morpheme "out" introduced into the sentence by some other means, the tree-fitting heuristic would not have to deal with the morpheme. As a putative universal for natural languages, the Graph Deformation Condition can be criticized in that the HAM structures representing the meanings of various sentence types are not specified a priori, but seem to be made up as they are needed. For this reason it is hard to disconfirm the Condition with the present examples, though.

Do natural languages permit Semantics-Induced Generalizations?

The Tree-fitting heuristic has a function other than giving a tree-structure to the sentence. The heuristic attaches semantic labels to the branches of the tree, and the Semantics-Induced Equivalence of Syntax heuristic uses these labels as criteria for merging rules derived from different sentences. These heuristics serve LAS well, but only because the subset of English grammar and the subset of HAM structures that Anderson has chosen correspond almost feature for feature. For example, the grammatical rule that specifies that sentences consist of a noun phrase and a verb phrase corresponds to the breakdown of HAM propositions into a subject and a predicate; the grammatical rule that breaks the predicate phrase into a spatial preposition and a noun phrase corresponds to the breakdown of a HAM predicate into a relation and an object, and so on. However, whenever syntax and semantics diverge, I will show, LAS errs, either over- or undergeneralizing.

Semantics-induced undergeneralizations

LAS's powers to generate more sentences than it has seen reside in its abilities to merge the different exemplars of a constituent type into a single class. Thus one would want LAS to recognize, say, that (to a first approximation) all noun phrases in English are generated by the same set of rules, regardless of the type of sentence or the position in a sentence in which the noun phrase is found. However LAS fails to do so even with the restricted subset of English it is given. For example, it fails to recognize the equivalence of subject noun phrases in sentences using the word "above" with those using the word "below". This is because the concepts "above" and "below" are represented identically in the propositions at the semantic level, with the subject of such a proposition interpreted by other procedures as the higher of the two objects in space. Thus the counterpart to "the square" in "the square is above the circle" is the subject in the underlying proposition, whereas in "the square is below the triangle" it is the object. For this reason the two occurrences of the phrase are mistakenly treated as different syntactic units.

Although Anderson suggests a solution to this particular problem, related problems will pop up when different subsets of languages are attempted. This is because natural languages frequently use the same constituents to express different underlying logical functions (which is one of the chief motivations for developing transformational grammars, with their distinction between deep and surface structures). Thus the Semantics-Induced Equivalence heuristic would never realize that the different tokens of the phrase "the cop" in 4(a-e)

4. (a) The cop frightens the thief.
   (b) The cop is frightened by the thief.
   (c) The cop tends to like thieves.
   (d) The cop who arrests thieves...
   (e) The cop who thieves frighten...

are examples of the same type of sentence constituent, since in the different sentences and phrases it functions variously as subject or object of the underlying proposition, or as part of the principal proposition or one of the secondary propositions. LAS would develop ad hoc rules for the different types of sentences, and would be unable to conclude that a subject noun phrase in an active sentence can also appear as the subject of a passive sentence, a "tend"-type sentence, and so on.

One interesting way to remedy this problem would be to posit distinct mental predicates corresponding to the different syntactic constructions that a verb can enter into. Thus there would be mental predicates for "FRIGHTEN", "IS-FRIGHTENED-BY", "TENDS-TO-FRIGHTEN", "IS-EASY-TO-FRIGHTEN", and so on (which is similar to a proposal Anderson has made elsewhere in discussing memory for sentences, see Anderson and Bower, 1973). Since the subjects of the sentences with all these constructions are also the subjects of their underlying propositions at the semantic level, LAS would have grounds to merge them. Unfortunately, this raises the problem of how the learner could tell when to encode a situation using one type of mental predicate rather than another. For example, how would the learner know to use the "FRIGHTEN" predicate just when hearing "It is easy to frighten the cat", but the "IS-EASY-TO-FRIGHTEN" predicate when hearing "That cat is easy to frighten"? This "encoding problem" and its possible solutions will be discussed further in Section IX.

Semantics-induced overgeneralizations

In relying on semantic criteria, LAS also generalizes in cases where it should not. For example, the proposition asserting that an entity is square-shaped can appear in a sentence either as "the square" or "the square thing", but the proposition asserting that an entity is colored red can appear only as "the red thing". Nonetheless, since the two propositions have the same format, LAS overgeneralizes and accepts "the red". Anderson solves this problem by providing LAS with an innate schema for noun phrases, including the stipulation that a noun phrase must contain at least one noun. If indeed the general form of the noun phrase is innate, it cannot have the format Anderson proposes, however, since many noun phrases lack nouns – consider the subject noun phrases in the sentences "Jogging exhausts me", "It is a total bore", and "That he chortles is irritating".

A similar overgeneralization problem follows from the fact that verbs with similar semantic representations have different case structures, that is, require different numbers and arrangements of noun phrases in the sentences in which they appear. Thus "give" might appear in a semantic structure with a subject and two objects, corresponding to the giver, gift, and recipient. Using this structure, LAS could build rules that parsed "Rockefeller gave Brown a million dollars", with two consecutive noun phrases after the verb; it would also construct rules to parse "Rockefeller gave a million dollars to Brown", with a noun phrase and a prepositional phrase. However when LAS later encounters sentences like "The IMF transferred a billion dollars to Ghana", or "Rockefeller donated a Wyeth to the museum", it would merge "give", "transfer", and "donate" into a single class, since they would have similar roles in their semantic representations, and would mistakenly produce "Rockefeller donated the museum a Wyeth", "The IMF transferred Ghana a billion dollars", and so on. Anderson does suggest a heuristic that might help LAS in learning the case structures of verbs: placing all the concepts that are causally related to the verb at the same level of embedding in the Prototype structure. This would not help for the present examples, however, since the different verbs have the same causal relation to the noun phrases, but have different case structures nonetheless. Many similar examples can be found in English: "throw" versus "propel", "show" versus "display", "teach" versus "instruct", and so on.

Learning grammatical morphemes

The class of grammatical morphemes (e.g., articles, inflections, conjunctions, relative pronouns, etc.) poses special problems for LAS, since they have no counterparts in its semantic structures. Anderson argues that learning the rules for ordering these terms in the absence of semantic information is *not* problematic, at least in a formal sense. Since grammatical morphemes occur in sub-sequences of finite length, they constitute a finite cardinality language, which, according to Gold's theorems, can be learned with neither an informant nor a semantic referent. This argument is misleading, however, because whether or not a string of grammatical morphemes is acceptable will depend on its context. Since the relevant context can be indefinitely long, there would be an infinite number of cases for the finite cardinality learner to memorize. Thus what the learner faces is *not* a finite cardinality language after all. For example, the occurrence of the string "to which" in sentence 5(a) is grammatical only because the verb "give", which can take a prepositional phrase beginning with "to", appears later in the sentence (compare the same sentence with "spent" in place of "gave"). But as sentences 5(b-d) show, that verb can be an arbitrary distance away, resulting in an infinite number of contexts to learn.

5. (a) The museum to which he gave a million dollars is in Chicago.
(b) The museum to which it is obvious he gave a million dollars is in Chicago.
(c) The museum to which I think it is obvious he gave a million dollars is in Chicago.

(d) The museum to which I think without any justification whatsoever it is obvious he gave a million dollars is in Chicago.

Thus learning rules for these classes of items is formally far from a trivial matter, and it is worth examining the heuristic solutions to the problem that Anderson proposes.

To begin with, it should be noted that LAS faces a language with few grammatical morphemes: only the articles "the" and "a", the copula "is", and the relative pronoun "which". The spatial prepositions such as "above" are treated as content words, since they correspond directly to nodes in the semantic representation, and to simplify matters even further, the expression "to the left of" has been collapsed into the single word "left-of". With this simple language, LAS can survive with a single heuristic: when it encounters one or more grammatical morphemes, it brackets them with the content word immediately to the right, creating a new constituent.

Problems with the grammatical morpheme heuristic

Although this heuristic works well enough to prevent LAS from making any gross errors, it prevents it from making important generalizations as well. For example, LAS cannot recognize the equivalence in its grammar of predicate phrases in the main clause of a sentence and predicate phrases in relative clauses, because the latter have the word "which" grafted onto them. This also seems to be the reason that LAS fails to merge its class for prenominal adjectives ("*red* square") with its identical class for predicate adjectives ("the square *is red*"). In any case, the heuristic clearly would not work for larger subsets of natural languages. As Anderson notes, in sentences like

6. The woman that he ran after is nimble.

LAS would create the nonsense constituent "after is nimble", leading to many possibilities for error (e.g., "The woman he loved after is nimble").

"Correcting procedures" for handling grammatical morphemes

Anderson does suggest remedies for some of these problems. For example, the first problem could be solved by allowing LAS to merge arcs with identical subconstituents, whether or not one arc is wholly contained in the other. However this procedure would still not make the required generalization in the general case – it would not help detect the similarities between main clauses and other sorts of relative clauses, such as those in which the objects have been deleted. For example, in 7(b), there is no constituent corresponding to the "the monster devoured" in 7(a), as the brackets indicate. Nonetheless one would want a learner to be able to generalize that whatever can be expressed in a main clause like 7(b)

7. (a) The cookie that the monster devoured is huge.
   (b) (The monster) (devoured (the cookie))

can also be expressed in a relative clause like the one in 7(a).

Anderson also suggests that redundant word classes, such as the predicate and prenominal adjective classes in our example, should be merged if they have enough members in common. But this would only lead to trouble. In natural languages, many if not most nouns can also serve as verbs and adjectives, but it would be disastrous to merge those classes outright, since many adjectives and verbs *cannot* serve as nouns.

Finally, Anderson suggests that the incorrect parse of sentences like 6 could be avoided if the learner would exploit the pause often found after the preposition in spoken speech as a cue to the correct location of the constituent boundary. However, natural speech is full of pauses that do *not* signal phrase boundaries (see Rochester, 1973), so such a heuristic would not, in general, do much good.

Conclusion

In sum, careful scrutiny of the learning mechanisms of LAS does not bear out Anderson's claim that such mechanisms are sufficient to learn natural languages. We have seen a number of cases in which the semantics-based heuristics are inadequate in principle to learn important features of English. This would not be a serious criticism if there were principled ways of extending LAS to handle these features. But virtually all of Anderson's proposals for extending LAS would at best work for the particular glitches they were designed to fix, and would be ineffective if applied to larger subsets of natural languages.

None of this diminishes the importance of Anderson's contribution. In the traditional psycholinguistic literature, the "Cognitive" theory of language learning is usually discussed in such vague terms that it is impossible to evaluate. In embodying this theory in a computer program, Anderson has shown what assumptions the theory rests on, which aspects of language learning the theory can account for, and which aspects are beyond its reach. In Section IX, I will discuss further the implications of LAS and other models for theories of human language learning.

VIII. A Theory of Learning Transformational Grammars

The features of natural language that give LAS the most trouble are precisely those features that cannot easily be handled by context-free grammars, and that motivated the development of transformational grammars (Chomsky, 1957, 1965). Examples are discontinuous constituents, "respectively"-type constructions, case- and complement structures of various verbs, the divergence of semantic roles and syntactic constituent structures, the placement of "grammatical morphemes", and generalizations that hold across related syntactic constructions. An adequate theory of language learning will have to account for the acquisition of languages with these sorts of properties. Henry Hamburger, Kenneth Wexler, and Peter Culicover have taken a large step in this direction by constructing a mathematical model which incorporates some reasonable assumptions about the language learner, and which they prove is capable of learning transformational grammars of a certain type (Hamburger and Wexler, 1975; Wexler, Culicover, and Hamburger, 1975; Culicover and Wexler, 1977).

Central to Hamburger, Wexler, and Culicover's theory is the assumption that the learner is innately constrained to entertain hypotheses of a certain sort, and is therefore capable of acquiring only certain types of languages. As I have mentioned, this assumption could conceivably enable an enumerative language learner to learn a language with access only to a sample of sentences. The assumption is also implicit in a weak form in the heuristic approach to language learning, and is explicitly embraced by Anderson when he claims that the learner "assumes" that the target language conforms to the Graph Deformation Condition and to the Semantics-Induced Equivalence of Syntax Condition. But Hamburger *et al.*, take the strongest view, originally proposed by Chomsky (1962, 1965), that innate, language-specific constraints cause the child to consider only a very narrowly-defined class of transformational grammars. Hamburger, Wexler, and Culicover's feat was to define these constraints in a precise way, show why they contribute to learnability, and make the case that natural languages fall into the class they define.

Hamburger *et al.*, begin with a version of Chomsky's transformational grammar, in which a set of context-free *base rules* generates a *deep structure* tree which *transformations* operate upon to produce a sentence. The base rules can generate arbitrarily large deep structures only by rewriting sentences within sentences, that is, by repetitively applying one of the rules that rewrites the "S" symbol. Each occurrence of an "S" delineates a *level* in the deep structure. Transformational rules are applied first at the lowest level (i.e., the most deeply embedded subsentence), then to the second lowest level, and so on.

Learnability of transformational grammars from a text

Wexler and Hamburger (1973) first attempted to prove that a constrained class of transformational grammars was identifiable in the limit from a sample of sentences (see the section on "Constraining the Hypothesis Space" in

Section IV). They made the assumption, known to be overly strong, that all languages have identical base rules and differ only in their transformational rules. Thus they made the base rules innate, and required the learner to identify in the limit a set of transformations that generated the target language. This they provec to be impossible. Therefore, in their next attempts (Hamburger and Wexler, 1975; Wexler, Culicover, and Hamburger, 1975) they assumed, with Anderson and the "Cognitive" theorists, that the child has simultaneous access to a string and its meaning, and must learn rules that translate one into the other.

Semantic representations and the Invariance Principle

In Hamburger *et al.*'s model, a sentence meaning is represented by a tree structure that has the same hierarchical breakdown of constituents as the deep structure of the sentence, but with no particular left-to-right ordering of the constituents (such a structure is similar to Anderson's "Prototype structure"). Since deep structure constituents are ordered differently in different languages, the first task for the learner is to learn the base rules which define the orderings his language uses. Wexler and Culicover note that this can be accomplished in a number of simple ways (in fact, Anderson's Tree-fitting heuristic is one such way). Like Anderson, they point out that this assumes that in all natural languages the deep structures *will* preserve the hierarchical connectivity of nodes in semantic structures, differing only in their linear order (i.e., branches may not cross, nor may links be severed and re-attached elsewhere). They justify this *Invariance Condition* (similar, of course, to Anderson's Graph Deformation Condition) by showing that out of all the combinatorial possibilities for ordering constituents of a certain type in deep structures, only those that respect the Invariance Condition are found in natural languages (over 200 of which they examine, Culicover and Wexler, 1974).

The learning procedure

From then on the learner must hypothesize a set of transformations, or a *transformational component*, that in combination with the base rules generates the target language. The procedure is simple. The learner undergoes an infinite series of trials in which he is presented with a meaning-sentence pair and is required to guess a grammar. For each pair, the learner applies his current transformational rules to the deep structure (which he computes from the meaning structure), and compares the result against the input string. If they match, the learner leaves his grammar untouched and proceeds to the next pair. If they do not match, the learner randomly decides between two courses of action. He can discard, at random, any of the transformations he used to derive the incorrect string; or, he can hypothesize a set consisting of all the transformations capable of transforming the deep structure to the input string in conjunction with the rest of the grammar, and select one of these transformations at random for inclusion in the grammar. Hamburger *et al.*, prove that with suitable constraints on the transformations used by the target language (and hypothesized by the learner), the learner will converge on a correct grammar for the language (i.e., the probability that the learner will have guessed a correct grammar becomes arbitrarily close to 1 as time passes). The proof is long and complex and will not be outlined here. Instead I will summarize how the constraints that Hamburger *et al.*, propose function to guarantee learnability. This, of course, is the crux of the Chomskian claim that learnability considerations favor a strongly nativist theory of language acquisition.

Proving learnability

As I have mentioned in Section IV, restricting the learner's hypothesis space only yields learnability if the intersection between the grammars in the hypothesis space and the grammars consistent with the sample becomes smaller and smaller as learning proceeds (see Figure 2). Hamburger *et al.* must show that when the learner has guessed an incorrect transformational component, he need not wait an arbitrarily long time before discovering his error, that is, encountering a semantic structure that the Component does not properly transform into the corresponding sentence. This in turn implies that the learner must not have to wait until an arbitrarily *complex* meaning-sentence pair appears in the sample before knowing that his transformational component is incorrect, since by the laws of probability he would have to wait an arbitrarily long time for an arbitrarily complex pair. In other words, if the learner has an incorrect transformational component, that component must make an error on a sentence-meaning pair that is no more complex than a certain bound (where complexity is measured by the number of S-nodes or levels in the deep structure).

This condition is not satisfied for unconstrained transformational grammars. In transformational grammars, each transformation is triggered by a particular configuration of symbols in a deep structure, or *structural description.* If a structural description can be arbitrarily complex for a transformation in the grammar, then the learner would have to wait until a meaning-sentence pair of that (arbitrary) complexity appeared in the sample before having occasion to hypothesize such a transformation. It would then be impossible to prove that the probability of the learner having hypothesized a complete, correct grammar approaches unity with increasing exposure to the sample. So Hamburger *et al.*, proposed the following constraint on transformations: no transformation may have a structural description that refers to symbols in more than two adjacent levels in the deep structure. Consider the deep structure-sentence pair in Figure 6 (the example has been simplified drastically from Chomsky, 1973). Assuming that the learner's transformational component does not yet correctly map one onto the other, the learner could hypothesize something like the following transformation (assuming that other transformations place the grammatical morphemes properly):

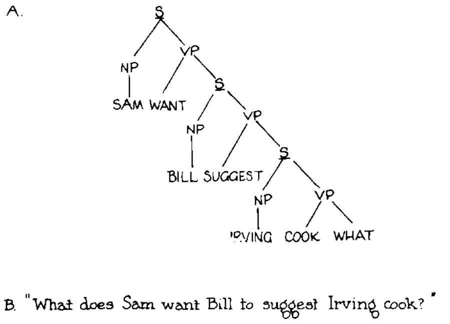
NP VP NP VP NP VP what → what NP VP NP VP NP VP.

However this transformation would be forbidden by Hamburger *et al.*'s constraint, because the symbols on the left hand side span across three levels in the deep structure. Instead, the learner could hypothesize something like the following:

NP VP what → what NP VP

which, applied successively from the deepest level upward, would produce the same string. (It is interesting to note that in this example the learner would not even have had to wait until encountering a pair this complex to hypothesize the transformation – an interrogative sentence with one level would have sufficed.) Hamburger *et al.*, argue that virtually all transformations in English and other languages conform to this condition, which they call the *Binary Principle.* Although they proposed the principle because, without it, they could not have proved learnability, they point out that Chomsky (1973) independently proposed an identical constraint, the *Subjacency Condition*, which he justified on descriptive grounds. That is, there seems to be independently motivated evidence that the Binary Principle is true of natural languages.

Figure 6. *Deep structure* (A) *and string* (B) *illustrating the Binary Principle.*



The Freezing Principle

The Binary Principle is not sufficient, however, to guarantee that an incorrect transformational component will make a telltale error on a meaning-sentence pair less complex than a certain bound. The base rules of a grammar can generate only a finite number of structures within a single level, by definition. Together with the Binary Principle, this would seem to ensure that input data of bounded complexity would suffice to exhaust all the structural descriptions that could trigger transformations. Unfortunately, whenever a transformation is applied at one level, it can alter the configuration of symbols within another level, creating new potential structural descriptions for transformations. Thus a series of transformations starting arbitrarily far down in a deep structure can alter the configuration of symbols within another level (as, in fact, the example in Figure 6 showed), creating new potential structural descriptions for transformations. A learner whose transformational component was in error only when applied to this altered configuration would never discover the error until coming across this arbitrarily complex structure. To remedy this situation, Culicover, Wexler, and Hamburger (1975) proposed a new constraint, the *Freezing Principle*, which forbids a transformation to apply to a configuration of symbols that could only have been created by the previous application of another transformation. The artificial example in Figure 7 shows how the constraint works. Say the learner must transform the deep structure 7(a) into the string 7(c), and already has a transformation that reverses the two morphemes C and B, as shown in 7(b). Now he must coin a transformation that reverses the morphemes A and B. The following transformation, for example, would accomplish this reversal:

ABDC → BADC.

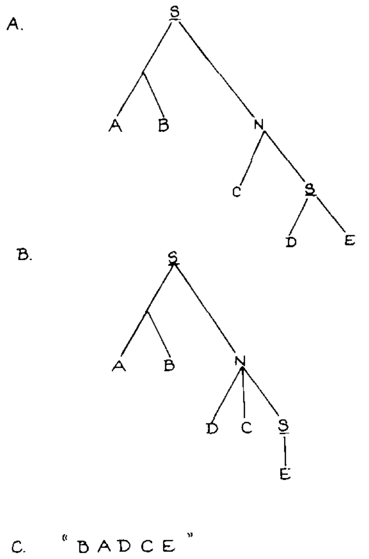
However, the Freezing Principle forbids this hypothesis, since it refers to the symbol sequence DC, which was not generated by a base rule but was created by another transformation. Instead, the learner can hypothesize the following transformation:[3]

AB → BA.

With the Binary and Freezing Principles, Hamburger, Wexler, and Culicover not only prove that the learner will converge on a correct grammar, but that

[3] As the example suggest, the Binary and Freezing Principles tend to reduce the context-sensitivity of rules in grammars by preventing large parts of tree structures from entering into the structural descriptions of transformations. This is not a coincidence, since in general context-free rules are more easily learnable than context-sensitive rules. See also Kaplan (1978) who argues that the reduction of context-sensitivity afforded by the Subjacency (i.e., Binary) Principle contributes to efficient sentence parsing.

Figure 7. *Deep structure* (A) *and string* (C) *illustrating the Freezing Principle.*



he can do so without even having to consider any structure with more than two levels of embedded sentences (i.e., three S-nodes).

Of course, Hamburger, Wexler, and Culicover must show that their constraint does not prevent their learner from acquiring any natural language. In Wexler, Culicover, and Hamburger (1975) and Culicover and Wexler (1977), examples of many sorts of English constructions are adduced to support the contention that natural languages obey the Freezing Principle. Moreover, Wexler *et al.* argue that in some cases the Freezing Principle does a better job than other constraints proposed in the linguistics literature at explaining why certain types of sentences are judged ungrammatical, and that in other cases, it mandates a choice between competing, hitherto equally plausible theories.

An evaluation

In evaluating the Hamburger *et al.*, model, it is important to note that I have changed the emphasis somewhat from their presentation. Their chief goal was to develop an "explanatorily adequate" linguistic theory (see Chomsky, 1965), which not only accounts for various linguistic phenomena, but *explains* why they must be one way and not another. Thus Hamburger, Wexler, and Culicover claim that the reason why natural languages conform to the Invariance, Binary, and Freezing Principles is that if they did not, they would not be learnable. Their model of a language learner was their means of justifying the claim.

Secondarily, they present their learning model as a first step toward an adequate theory of language learning (which is what I have been emphasizing). As such, they can claim no more than that their model is (at least) "minimally plausible". It requires no information about non-sentences, does not have to remember the entire sample, requires sentences no more complex than those with two levels of subsentences, employs semantic information in learning, processes sentences one at a time, and changes its gramma rule by rule. In other words, it does not flagrantly contradict some of the obvious facts of human language development. However, since the model is only a specification of the boundary conditions of a theory of language learning (i.e., they are claiming that the child's hypotheses must be *no less constrained* than those of the model), many features would have to be fleshed out before it could be considered any more than "minimally plausible". First, there is no indication at present that the learner would converge in a time-span comparable to a human childhood. It seems inefficient and implausible to have the child enumerating sets of transformations and mentally rolling dice to decide which to keep or discard. What is needed is a theory showing how the child's hypotheses are guided in a more direct way by the meaning-sentence pair under consideration, and how these hypotheses are computed during the left-to-right processing of a sentence. Third, unordered deep structures are questionable candidates for a theory of the child's representational system (although this will be discussed further in Section IX). Finally, we are left with few suggestions as to how the transformational component, once acquired, is used in producing and comprehending sentences.

In any case the Hamburger, Wexler, and Culicover model is a unique and extremely impressive achievement. Theirs is the only model that is capable of learning natural languages in all their complexity and that at the same time is not blatantly counter to what we know about the child and his learning environment. They also have clarified and justified, more clearly than anyone else has, two central tenets of transformational linguistics: that considerations of language learnability can dictate a choice between rival linguistic theories, and that learnability considerations imply strong innate constraints of a certain type on the child's language learning faculties. As they put it, "The bridge that Chomsky has re-erected between psychology and linguistics bears two-way traffic" (Hamburger and Wexler, 1975).

IX. Implications for Developmental Psycholinguistics

Toward a Theory of Language Learning

Among the models of language learning that I have considered, two seem worthy upon examination to serve as prototypes for a *theory* of human language acquisition. Anderson's LAS program roughly meets the Cognitive, Input, and Time Conditions, while faring less well with the Learnability and Equipotentiality Conditions. Hamburger, Wexler, and Culicover's transformational model meets the Learnability and Equipotentiality Conditions (clearly), and the Input Condition (perhaps), while faring less well with the Cognitive and Time Conditions. I hope it is not too banal to suggest that we need a theory that combines the best features of both models. It must incorporate a psychologically realistic comprehension process, like Anderson's system, since language acquisition is most plausibly thought of as being driven by the comprehension process. But at the same time, the model's semantic structures must be rich enough, and the hypothesization procedure constrained enough, that any natural language can be shown to be learnable (like the Hamburger *et al.*, model), so that the model does not become buried under a pile of ad hoc, semi-successful heuristics when it is extended to more and more linguistic domains. Of course, developing such a theory has been hampered by the lack of a suitable theory of language itself, one that both gives a principled explanation for linguistic phenomena in various domains and languages, and that can be incorporated in a reasonable way into a comprehension model (see Bresnan, 1978, for a step in this direction). Of course, here is not the place to attempt to present a new theory synthesizing the best features of previous efforts. Instead, I will attempt to point out the implications that the formal study of language learning has for current issues in developmental psycholinguistics.

Developmental Psycholinguistics and Language Acquisition Devices

Current attitudes toward language acquisition models among developmental psycholinguists have been strongly influenced by the fate of a research framework adopted during the 1960's that went under the name of the Language Acquisition Device, or LAD. There were in fact two different meanings to the expression Language Acquisition Device, and I think it is important to distinguish them. In one formulation (Chomsky, 1962), the child was idealized as an abstract device that constructed rules for an unknown language on the basis of a sample of sentences from that language; characterizing the workings of that "device" was proposed as a goal for linguistics and psychology. As an analogy, we could think of a physiologist interested in electrolyte regulation who idealized the brain as "a bag of salt water", proceeding then to study the structure of the membrane, concentration of ions, and so on. Of course, in this sense, I have been talking about language acquisition devices throughout the present paper. However there is a second, stronger sense in which LAD is taken to describe a specific theory of language acquisition (e.g., Clark, 1973; Levelt, 1973). In this sense (Fodor, 1966; McNeill, 1966), the child is said to possess an innate mental faculty containing highly specific knowledge about transformational grammars, which extracts deep structures from the speech around him and adopts transformational rules, one at a time, culminating in a transformational grammar for the language. Pursuing the analogy with physiology, LAD would correspond in this sense to our physiologist proposing that the brain accomplished electrolyte regulation by means of a special purpose structure, "a bag of salt water", with various properties. In support of this theory, it was claimed that the child based his learning on a sample of speech composed largely of fragments and complex, semi-grammatical expressions (Chomsky, 1965), that the early utterances of the child displayed mastery of highly abstract syntactic relations (McNeill, 1966), and that the linguistic progress of the child seemed to reflect the accretion of transformations (e.g., Brown and Hanlon, 1970). However the entire approach quickly fell into disfavor when it was found that the speech directed to children was well-formed and structurally simple (Snow, 1972), that the child might exploit semantic information in addition to sentences themselves (e.g., Macnamara, 1972), that the early speech of children might be better broken down into "cognitive" or semantic relations than into abstract syntactic ones (e.g., Bowerman, 1973; Brown, 1973), and that in many cases children learned transformationally complex constructions before they learned their simpler counterparts (e.g., Maratsos, 1978). As a result, LAD has been abandoned by developmental psycholinguists as a *theory*, and in its place I think there has developed a rough consensus that semantic and pragmatic information, together with the simplified speech of parents, allows children to learn language by using general cognitive skills rather than a special language-specific faculty. However, LAD has also been rejected in its more general sense as a *problem* to be addressed, and it also seems to me that most debates in developmental psycholinguistics are, unfortunately, no longer carried out with an eye toward ultimately specifying the mechanisms of syntax acquisition. When specific proposals concerning such mechanisms are considered, I shall argue, the substance of many of these debates can change significantly.

Nativism versus Empiricism: Two Extreme Proposals

Formal results from the study of language learnability give us grounds for dismissing quite decisively two general proposals concerning what sort of mechanisms are necessary and sufficient for language learning, one empiricist, one nativist.

The extreme empiricist proposal is that there are no language-specific a priori constraints on the types of rules that humans can acquire. In this vein, it is argued that once a sufficient number of sentences has been observed, languages can be learned by "general multipurpose learning strategies" (Putnam, 1971), by "discovery procedures" (Braine, 1971), or by "learning algorithms" like a "discretizer-plus-generalizer" that "extracts regularity from the environment" (Derwing, 1973). As I have mentioned, Gold's enumeration procedure is the most powerful imaginable realization of a general learning algorithm. Nevertheless, even this procedure is inadequate in principle to acquire rules on the basis of a sample of sentences. And if the criterion for "acquisition" is weakened (by requiring only approachability, approximations to the target language, etc.), then learning is possible, but not within a human lifespan.

At the other extreme is the proposal that innate knowledge of the properties of natural languages, especially those of deep structures, allows the child to learn a language from a sample of sentences (e.g., Fodor, 1966; McNeill, 1966). In one of Hamburger and Wexler's early models (Wexler and Hamburger, 1973), they imposed constraints on the learner's hypotheses that were known to be unrealistically stringent (e.g., that all languages share identical deep structure rules). Nevertheless they proved that this class of languages is unlearnable on the basis of a sample of sentences, and therefore, that the same must be true of classes that are specified more weakly (and hence more realistically). Of course, it is still possible that a different sort of innate constraint might guarantee learnability, but this will remain a matter of speculation until someone puts forth such a proposal.

Problems for the Cognitive Theory of Language Acquisition

The inability of these procedures to induce grammars from samples of sentences suggests strongly that semantic and pragmatic information is used in language learning. The moderate success of the models of Anderson and of Hamburger *et al.*, also lends credence to this conclusion. However, despite the great popularity of the Cognitive Theory among developmental psycholinguists, there has been little discussion of what I believe to be the foundation of the theory: the precise nature of the child's internal representations. The Cognitive Theory requires that children have available to them a system of representational structures similar enough in format to syntactic structures to promote language learning, and at the same time, flexible and general enough to be computable by children's cognitive and perceptual faculties on the basis of nonlinguistic information. Until we have a theory of the child's mental representations that meets these conditions, the Cognitive Theory will remain an unsupported hypothesis. Unfortunately, designing a representational system with the desired properties will be far from a simple task. The two main problems, which I call the "encoding problem" and the "format problem", pit the Cognitive Condition against the Learnability and Equipotentiality Conditions.

The encoding problem

This problem is a consequence of the fact that languages can describe a situation in a number of ways, and that humans can perceive a situation in a number of ways. One might plausibly attribute many different representational structures to a child perceiving a given situation, but only one of these structures will be the appropriate one to try to convert into the sentence being heard simultaneously. Barring telepathy, how does the child manage to encode a situation into just the structure that underlies the sentence that the adult is uttering?

Consider an earlier example. Anderson assumes that when a child sees, say, a white cat eating a mouse, his mind constructs a structure something like the one in Figure 3(a). This is fortunate for the child (and for the model-builder), since in the example the sentence arriving concurrently happens to be "The white cat eats a mouse", whose meaning corresponds to that structure. But what if the sentence were "The mouse is being eaten by the cat", "That's the second mouse that the cat has eaten", "Some cats don't eat mice", "What's that white cat doing with the mouse?", and so on? To put it differently, assuming that the original sentence was the one uttered, what if the child were to have constructed a cognitive structure containing propositions asserting that the mouse was 'all gone', or that the cat and mouse were playing, or that the mouse looked easy for the cat to eat, and so on? In any of these cases, the child would face the task of trying to map a meaning structure onto a string with which it has only a tenuous connection. Thus the semantic representation would offer few clues, or misleading ones, about how to hypothesize new rules.[4]

[4] Dan Slobin (1978; personal communication) has pointed out that the child faces a similar problem in learning the morphology of his language. Natural languages dictate that certain semantic features of the sentence referent (e.g. number, person, gender, definiteness, animacy, nearness to the speaker, completedness, and so on) must be signalled in prefixes, suffices, alternate vowel forms, and other means. However, these features are by no means all that a child *could* encode about an event: the color, absolute position, and texture of an object, the time of day, the temperature, and so on, though certainly perceptible to the child, are ignored by the morphology of languages, and hence should not be encoded as part of the semantic structure that the child must learn to map onto the string. To make matters worse, the morphological rules of different languages select different subsets of these features *(continued opposite)*

I have already mentioned that Anderson would face this problem if he were to multiply the number of available mental predicates that correspond to a given verb, in order to foster certain generalizations. Hamburger *et al.* face a similar problem. In their model, the structures underlying synonymous sentences, such as actives and passives, are presumably identical except for a marker triggering a transformation in cases like the passive (since each transformation is obligatorily triggered by some deep structure configuration). Again, it is not clear how the child knows when to insert into his semantic structure the markers that signal the transformations that the adult happens to have applied.

Possible solutions to the encoding problem

I see three partial solutions to the encoding problem that together would serve to reduce the uncertainty associated with typical language learning situations, ensuring that the child will encode situations into unique representations appropriate to the sentences the adult is uttering. The first relies on the hypothesis that the representational system of the child is less powerful and flexible than that of the adult, and is capable of representing a given situation in only a small number of ways. Thus in the preceding example, the child is unlikely to encode the scene as propositions asserting that the mouse was not eating the cat, that all cats eat mice, etc. As the child develops, presumably his representational powers increase gradually, and so does the range of syntactic constructions addressed to him by his parents. If, as is often suggested (e.g., Cross, 1977), parents "fine-tune" their speech to the cognitive abilities of their children, that is, they use syntactic constructions whose semantics correspond to the representations most likely to be used by the child at a given moment, then the correspondence between the adult's sentence meaning and the child's encoding of the situation would be closer than we have supposed.

The second solution would posit that the child's social perception is acute enough to detect all the pragmatic or communicative differences that are concurrently signaled by syntactic means in different sentences (see Bruner, 1975). That is, the child knows from the conversational context what the adult is presupposing, what he or she is calling attention to, what is being asserted of what, and so on. For example, the child must not only *see* that

to signal obligatorily, and disagree further over which features should be mapped one-to-one onto morphological markers, and which sets of features should be conflated in a many-to-one fashion in particular markers. Thus there has to be some mechanism in the child's rule-hypothesization faculty whereby his possible conceptualizations of an event are narrowed down to only those semantic features that languages signal, and ultimately, down to only those semantic features that his target language signals.

the cat is eating the mouse, but must know that the adult is asserting of the cat that it is eating a mouse, instead of asserting of the mouse that it is disappearing into the cat, or many other possibilities. (As mentioned earlier, Anderson used this rationale in developing LAS, when he marked one of the propositions in each semantic structure as the intended "main proposition" of the sentence.) If this line of reasoning is correct, strong conditions are imposed both on the language and on the learner. The syntax of languages must not allow synonymy, in a strict sense: any two "base" structures (i.e., Anderson's Prototype structure or Hamburger *et al.*'s deep structure) that do not differ semantically (i.e., instantiate the same propositions) must differ pragmatically in some way. Conversely, the pragmatic and perceptual faculties of the child must be capable of discriminating the types of situations that occasion the use of different syntactic devices.

The third solution would equip the child with a strategy that exploited some simple property of the sentence to narrow down the possible interpretations of what the adult is asserting. Anderson implicated a strategy of this sort when LAS examined the set of words in a sentence and retained only the propositions in its meaning structure whose concepts corresponded to those words. In the present example, the child might always construct a proposition whose subject corresponds to the first noun in the sentence, and then choose (or, if necessary, create) some mental predicate that both corresponds to the verb and is consistent with his perception of the scene. Thus, when hearing an active sentence, the child would construct a proposition with the cat as the subject and "EATS" as part of the predicate; when hearing the passive version, the proposition would have the mouse as the subject and "IS-EATEN-BY" as part of the predicate.[5] One can even speculate that such a strategy is responsible for Bever's (1970) classic finding that children of a certain age interpret the referent of the first noun of both active and passive sentences as the agent of the action designated by the verb. The children may have set up the concept corresponding to the first noun as the subject of a proposition, but, lacking mental predicates like "IS-EATEN-BY" at that stage in their development, they may have mistakenly chosen predicates like "EATS" by default.

I hope to have shown how consideration of the requirements and implications of formal theories of language learning (in this case, those of Anderson and of Hamburger *et al.*) lead one to assign more precise roles to several phenomena studied intensively by developmental psycholinguists. Specific-

[5] This example follows the Anderson model with the "multiple predicate" modification I suggested. In the Hamburger *et al.* model, the child could insert a "transformation marker" into his deep structure whenever the subject of the deep structure proposition was not the first noun in the sentence.

ally, I suggest that the primary role in syntax learning of cognitive development, "fine-tuning" of adult speech to children learning language, knowledge of the pragmatics of a situation, and perceptual strategies is to ensure that the child encodes a situation into the same representational structure that underlies the sentence that the adult is uttering concurrently (cf. Bruner, 1975; Bever, 1970; Sinclair de-Zwart, 1969; and Snow, 1972; for different interpretations of the respective phenomena).

The Format Problem

Once we are satisfied that the child has encoded the situation into a unique representation, corresponding to the meaning of the adult's sentence, we must ensure that that representation is of the appropriate *format* to support the structural analyses and generalizations required by the learning process.

To take an extreme example of the problem, imagine that the study of perceptual and cognitive development forced us to conclude that the internal representations of the child were simply lists of perceptual features. Using a semantics-based generalization heuristic, the learner would have no trouble merging words like "cat" and "mouse", since both are objects, furry, animate, four-legged, etc. But the learner would be unable to admit into this class nouns like "flutter" or "clang", which have no perceptual features in common with "cat", nor "fallacy" or "realization", which have no perceptual features at all. The difficulties would intensify with more abstract syntactic structures, since there are no conjunctions of perceptual features that correspond to noun phrases, relative clauses, and so on. The problem with this representational format is that even if it were adequate for perception, it is not adaptable to syntax learning. It does not provide the units that indicate how to break a sentence into its correct units, and to generalize to similar units across different sentences.

In other words, what is needed is a theory of representations whose elements correspond more closely to the elements of a grammar. In Anderson's theory, for example, a representation is composed of a "subject" and a "predicate", which in turn is composed of a "relation" and an "object". These correspond nicely to the syntactic rules that break down a sentence into a noun phrase and a verb phrase, then the verb phrase into a verb and another noun phrase. Furthermore, propositions encoded for different situations in which syntactically similar sentences would be uttered would all have the same format, regardless of whether they represent furry things, square things, events, actions, abstract mathematical concepts, or other propositions. Hamburger *et al.*, posit a cognitive representation with a format even more suitable to language learning: unordered deep structures. This is one of the reasons why their model is more successful at acquiring syntactic rules than

LAS is. In sum, these theorists posit that the syntax of the language of thought is similar to the syntax of natural languages.

However, this solution might create problems of its own. It is possible for theorists to use "cognitive" representations with a format so suitable to syntactic rule learning that the representations may no longer be plausible in a theory of perception or cognition. To take a hypothetical example, in standard transformational grammars a coordinated sentence such as "Jim put mustard and relish on his hot dog" is derived from a two-part deep structure, with trees corresponding to the propositions "Jim put mustard on his hot dog" and "Jim put relish on his hot dog". However a theory of cognitive or perceptual representations based on independent evidence (e.g., reaction times, recall probabilities, etc.), when applied to this situation, might not call for two separate propositions, but for a single proposition in which one of the arguments was divided into two parts, corresponding to the two conjoined nouns (which is the way it is done in Anderson and Bower, 1973, for example). Cases like this, if widespread and convincing, would undermine Hamburger *et al.*'s premise that unordered deep structures are plausible as cognitive representations.

In this vein, it is noteworthy that even though Anderson's semantic structures were lifted from his theory of long term memory, they too are more similar to linguistic deep structures than those of any other theory of memory representation, incorporating features like a binary subject-predicate division, distinct labels for each proposition, and a hierarchical arrangement of nodes (cf., Norman and Rumelhart, 1975; Winston, 1975). In fact, many of these features are not particularly well-supported by empirical evidence (see Anderson, 1976), and others may be deficient on other grounds (see Woods, 1975). Concerning other computer models in which "the designer feeds in what he thinks are the semantic representations of utterances", McMaster *et al.* (1976, p. 377) remark that "the risk is that [the designer] will define semantics in such a way that it is hardly different from syntax. He is actually providing high-level syntactic information. This gives the grammar-inferrer an easy task, but makes the process less realistic...".[6]

Implications of the format problem

Faced with possibly conflicting demands on a theory of the form of mental representation from the study of language learning and the study of other

[6] This discussion has assumed that the language-specific structures posited as cognitive representations are specific to languages in general, not to particular languages. If the representations are tailored to *one* language (e.g., when predicates in LAS's propositions take the same number of arguments as the verb they correspond to, even though the same verbs in different languages take different numbers of arguments), a second and equally serious problem results.

cognitive processes, we have two options. One is to assert that, all other considerations notwithstanding, the format of mental representations *must* be similar to syntactic structures, in order to make language learning possible. Fodor (1976), for example, has put forth this argument.[7] The second is to posit at least two representational formats, one that is optimally suited for perception and cognition, and one that is optimally suited for language learning, together with a conversion procedure that transforms a representation from the former to the latter format during language learning. Anderson and Hamburger *et al.*, already incorporate a version of this hypothesis. In LAS, the semantic structures are not entirely suitable for rule learning, so there is a procedure that converts them into the "prototype structures". And in the Hamburger *et al.*, model, the deep structures are not entirely suitable as cognitive representations (being too specific to particular languages), so there is a procedure whereby they are derived from "semantic structures". Ultimately the Cognitive Theory of language learning must posit one or more representational formats appropriate to cognition in general and to language learning in particular, and, if necessary, the procedures that transform one sort of representation into the other.

Nativism and empiricism revisited

It is often supposed that if children indeed base their rule learning on cognitive representational structures, the traditional case for nativism has been weakened (e.g., Schlesinger, 1971; Sinclair de-Zwart, 1969). According to this reasoning, cognitive structures already exist for other purposes, such as perception, reasoning, memory, and so forth, so there is no need to claim that humans possess an innate set of mental structures specific to language. However, this conclusion is at best premature. It is far from obvious that the type of representational structures motivated by a theory of perception or memory is suitably adaptable to the task of syntactic rule learning. For if the foregoing discussion is correct, the requirements of language learning dictate that cognitive structures are either language-like themselves, or an innate procedure transforms them into structures that are language-like. When one considers as well the proposed innate constraints on how these structures enter into the rule hypothesization process (i.e., Anderson's Graph Deformation and Semantics-Induced Equivalence Principles, and Hamburger *et al.*'s Binary and Freezing Principles), one must conclude that the Cognitive Theory

[7] Incidentally, it is ironic that Anderson, in a different context, fails to mention this argument when he examines the case for propositional theories of mental representation in general (Anderson, 1978).

of language learning, in its most successful implementations, vindicates Chomsky's innateness hypothesis if it bears on it at all.[8]

Language learning and other forms of learning

It might be conjectured that if one were to build models of other instances of human induction (e.g., visual concept learning, observational learning of behavior patterns, or scientific induction), one would be forced to propose innate constraints identical to those proposed by the designers of language learning models. If so, it could be argued that the constraints on language learning are necessitated by the requirements of *induction* in general, and not natural language induction in particular. While it is still too early to evaluate this claim, the computer models of other types of induction that have appeared thus far do not seem to support it. In each case, the representational structures in which data and hypotheses are couched are innately tailored to the requirements of the *particular domain of rules being induced.* Consider Winston's (1975) famous program, which was designed to induce classes of block-structures, such as arches and tables, upon observing exemplars and non-exemplars of the classes. The units of the program's propositional structures can designate either individual blocks, blocks of triangular or rectangular shape, or any block whatsoever; the connecting terms can refer to a few spatial relations (e.g., adjacency, support, contact) and a few logical relations (e.g., part-whole, subset-superset). The program literally cannot conceive of distance, angle, color, number, other shapes, disjunction, or implication. This removes the danger of the program entertaining hypotheses other than the ones the programmer is trying to teach it. Similarly. Soloway and Riseman's (1977) program for inducing the rules of baseball upon observing sample plays is fitted with innate knowledge of the kind of rules and activities found in competitive sports in general. Langley's (1977) program for inducing physical laws upon observing the behavior of moving bodies is confined to considering assertions about the values of parameters for the positions, velocities, and accelerations of bodies, and is deliberately fed only those attributes of bodies that are significant in the particular mock universe in which it is "placed" for a given run. These restrictions are not just adventitious shortcuts, of course. Induction has been called "scandalous" because any finite set of observations supports an intractably large number of gener-

[8] One could contest this conclusion by pointing out that it has only been shown that the various nativist assumptions are *sufficient* for learnability, not that they are *necessary*. But as Hamburger and Wexler put it (1975), "anyone who thinks the assumption[s are] not necessary is welcome to try to devise proofs corresponding to ours without depending on [those] assumptions".

alizations. Constraining the type of generalizations that the inducer is allowed to consider in a particular task is one way to defuse the scandal.

Parental Speech to Children

Frequently it is argued that the special properties of parents' speech to children learning language reduces the need for innate constraints on the learning process (e.g., Snow, 1972). Since these claims have not been accompanied by discussions of specific learning mechanisms that benefit from the special speech, they seem to be based on the assumption that something in the formal properties of the language learning task makes short, simple, grammatical, redundant sentences optimal for rule learning. However a glance at the models considered in the present paper belies this assumption: the different models in fact impose very different requirements on their input.

Consider the effects of interspersing a few ungrammatical strings among the sample sentences. Gold's enumeration learner would fail miserably if a malformed string appeared in the sample – it would jettison its correct hypothesis, never to recover it, and would proceed to change its mind an infinite number of times. On the other hand, Horning's Bayesian learner can easily tolerate a noisy sample, because here the sample does not mandate the wholesale acceptance or rejection of grammars, but a selection from among them of the one with the highest posterior probability. The Hamburger *et al.*, model would also converge despite the occasional incorrect input datum, since at any point in the learning process at which it has an incorrect grammar (e.g., if it were led astray by a bad string), there is a nonzero probability that it will hypothesize a correct grammar within a certain number of trials (assuming, of course, that it does not encounter another bad string before converging).

Similarly, it is doubtful that the length or complexity of sentences has a uniform effect on different models. Feldman described a procedure requiring that the sample sentences be ordered approximately by increasing length, whereas Gold's procedure is completely indifferent to length. In the Hamburger *et al*, model, contrary to the intuition of some, learning is facilitated by *complex* sentences – not only will the learner fail to converge if he does not receive sentences with at least two levels of embedded sentences, but he will converge faster with increasingly complex sentences, since in a complex sentence there are more opportunities for incorrect transformations or the absence of correct transformations to manifest themselves by generating the wrong string. Nevertheless, short and simple sentences may indeed facilitate learning in humans, but for a different reason. Since children have limited attention and memory spans, they are more likely to retain a short string of words for sufficient time to process it than they would a long string of words. Similarly, they are more likely to encode successfully a simple conceptualization of an event than a complex one. Thus short, simple sentences may set the stage for rule hypothesization while playing no role (or a detrimental role) in the hypothesization process itself.

Other models are sensitive to other features of the input. Since Klein and Kuppin's Autoling relies on distributional analysis, it thrives on sets of minimally-contrasting sentences. Since Anderson's LAS merges constituents with the same semantic counterparts, it progresses with sets of sentences with similar or overlapping propositional structures.

In sum, the utility of various aspects of the input available to a language learner depends entirely on the learning procedure he uses. A claim that some feature of parental speech facilitates rule learning is completely groundless unless its proponent specifies some learning mechanism.

Conclusions

In an address called "Word from the Language Acquisition Front", Roger Brown (1977) has cautioned:

"Developmental psycholinguistics has enjoyed an enormous growth in research popularity... which, strange to say, may come to nothing. There have been greater research enthusiasms than this in psychology: Clark Hull's principles of behavior, the study of the Authoritarian personality, and, of course, Dissonance Theory. And in all these cases, very little advance in knowledge took place. ...A danger in great research activity which we have not yet surmounted, but which we may surmount, is that a large quantity of frequently conflicting theory and data can become cognitively ugly and so repellent as to be swiftly deserted, its issues unresolved."

It is my belief that one way to surmount this danger is to frame issues in the context of precise models of the language learning process, following the lead of other branches of the cognitive sciences. I hope to have shown in this section why it may be necessary to find out how language learning *could* work in order for the developmental data to tell us how it *does* work.

References

Anderson, J. (1974) Language acquisition by computer and child. (Human Performance Center Technical Report No. 55.) Ann Arbor, University of Michigan.

Anderson, J. (1975) Computer simulation of a Language Acquisition System: A first report. In R. Solso (ed.), *Information processing and cognition: The Loyola Symposium*. Washington, Erlbaum.

Anderson, J. (1976) *Language, Memory, and Thought.* Hillsdale, N.J.: Erlbaum.

Anderson, J. (1977) Induction of augmented transition networks. *Cog. Sci., 1,* 125–157.

Anderson, J. (1978) Arguments concerning representations for mental imagery. *Psychol. Rev., 85,* 249–277.

Anderson, J. and G. Bower (1973) *Human Associative Memory.* Washington, Winston.

Bever, T. (1970) The cognitive basis for linguistic structures. In J. Hayes (ed.), *Cognition and the Development of Language.* New York, Wiley.

Biermann, A. and J. Feldman (1972) A survey of results in grammatical inference. In S. Watanabe (ed.), *Frontiers in Pattern Recognition.* New York, Academic Press.

Bowerman, M. (1973) *Learning to talk: A Cross-sectional Study of Early Syntactic Development, with Special Reference to Finnish.* Cambridge, U.K., Cambridge University Press.

Braine, M. (1963) The ontogeny of English phrase structure: The first phrase. *Lang., 39,* 1–14.

Braine, M. (1971) On two models of the internalization of grammars. In D. Slobin (ed.), *The Ontogenesis of Grammar.* New York, Academic Press.

Bresnan, J. (1978) A realistic transformational grammar. In G. Miller, J. Bresnan and M. Halle (eds.), *Linguistic Theory and Psychological Reality.* Cambridge, Mass., MIT Press.

Brown, R. (1973) *A First Language: The Early Stages.* Cambridge, Mass., Harvard University Press.

Brown, R. (1977) Word from the language acquisition front. Invited address at the meeting of the Eastern Psychological Association, Boston.

Brown, R, C. Cazden and U. Bellugi (1969) The child's grammar from I to III. In J. Hill (ed.), *Minnesota Symposium on Child Psychology, Vol. II.* Minneapolis, University of Minnesota Press.

Brown, R. and C. Hanlon (1970) Derivational complexity and order of acquisition in child speech. In J. Hayes (ed.), *Cognition and the Development of Language.* New York, Wiley.

Bruner, J. (1975) The ontogenesis of speech acts. *J. child Lang., 2,* 1–19.

Chomsky, N. (1957) *Syntactic Structures.* The Hague, Mouton.

Chomsky, N. (1962) Explanatory models in linguistics. In E. Nagel and P. Suppes (eds.), *Logic, Methodology, and Philosophy of Science.* Stanford, Stanford University Press.

Chomsky, N. (1965) *Aspects of the Theory of Syntax.* Cambridge, Mass., MIT Press.

Chomsky, N. (1973) Conditions on transformations. In S. Anderson and P. Kiparsky (eds.), *A Festschrift for Morris Halle.* New York, Holt, Rinehart and Winston.

Clark, E. (1973) What should LAD look like? Some comments on Levelt. In *The Role of Grammar in Interdisciplinary Linguistic Research.* Colloquium at the University of Bielefeld, Bielefeld, W. Germany.

Cross, T. (1977) Mothers' speech adjustments: The contribution of selected child listener variables. In C. Snow and C. Ferguson (eds.), *Talking to Children: Input and Acquisition.* New York, Cambridge University Press.

Culicover, P. and K. Wexler (1974) The Invariance Principle and universals of grammar. (Social Science Working Paper No. 55.) Irvine, Cal., University of California.

Culicover, P. and K. Wexler (1977) Some syntactic implications of a theory of language learnability. In P. Culicover, T. Wasow, and A. Akmajian (eds.), *Formal Syntax.* New York, Academic Press.

Derwing, B. (1973) *Transformational Grammar as a Theory of Language Acquisition.* Cambridge, UK, Cambridge University Press.

Fabens, W. and D. Smith (1975) A model of language acquisition using a conceptual base. (Technical Report CBM-TR-55, Department of Computer Science.) New Brunswick, N.J.) Rutgers – The State University.

Feldman, J. (1972) Some decidability results on grammatical inference and complexity. *Information and Control, 20,* 244–262.

Fodor, J. (1966) How to learn to talk: Some simple ways. In F. Smith and G. Miller (eds.), *The Genesis of Language.* Cambridge, Mass., MIT Press.

Fodor, J. (1975) *The Language of Thought.* New York, Thomas Crowell.

Fu, K. and T. Booth (1975) Grammatical inference: Introduction and survey. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-5(1),* 95–111; *SMC-5(4),* 409–423.

Gold, E. (1967) Language identification in the limit. *Information and Control, 16,* 447–474.

Gross, M. (1972) *Mathematical models in linguistics.* Englewood Cliffs, N.J., Prentice-Hall.

Hamburger, H. and K. Wexler (1975) A mathematical theory of learning transformational grammar. *J. Math. Psychol., 12,* 137–177.

Harris, Z. (1964) Distributional structure. In J. Fodor and J. Katz (eds.), *The Structure of Language.* Englewood Cliffs, N.J., Prentice Hall.

Hopcroft, J. and J. Ullman (1969) *Formal languages and their relation to automata.* Reading, Mass., Addison Wesley.

Horning, J. (1969) A study of grammatical inference. (Technical Report No. CS 139, Computer Science Dept.) Stanford, Stanford University.

Kaplan, R. (1975) On process models for sentence analysis. In D. Norman and D. Rumelhart (eds.), *Explorations in cognition.* San Francisco, W. H. Freeman.

Kaplan, R. (1978) Computational resources and linguistic theory. Paper presented at the Second Theoretical Issues in Natural Language Processing Conference, Urbana, Ill.

Kelley, K. (1967) Early syntactic acquisition. (Report No. P-3719.) Santa Monica, Cal., The Rand Corporation.

Klein, S. (1976) Automatic inference of semantic deep structure rules in generative semantic grammars. In A. Zampoli (ed.), *Computational and Mathematical Linguistics: Proceedings of 1973 International Conference on Computational Linguistics, Pisa.* Florence, Italy, Olschki.

Klein, S. and M. Kuppin (1970) An interactive program for learning transformational grammars. *Computer Studies in the Humanities and Verbal Behavior, III,* 144–162.

Klein, S. and V. Rozencvejg (1974) A computer model for the ontogeny of pidgin and creole languages. (Tehcnical Report No. 238, Computer Science Dept.) Madison: University of Wisconsin.

Knobe, B. and K. Knobe (1977) A method for inferring context-free grammars. *Information and Control, 31,* 129–146.

Kosslyn, S. and S. Schwartz (1977) A simulation of visual imagery. *Cog. Sci., 1,* 265–296.

Langley, P. (1977) BACON: A production system that discovers empirical laws. (CIP Working Paper No. 360.) Pittsburg, Carnegie Mellon University.

Levelt, W (1973) Grammatical inference and theories of language acquisition. In *The role of Grammar in Interdisciplinary Linguistic Research.*, Colloquium at the University of Bielefeld, Bielefeld, W. Germany.

Macnamara, J. (1972) Cognitive basis for language learning in infants. *Psychol. Rev., 79,* 1–13.

Maratsos, M. (1978) New models in linguistics and language acquisition. In G. Miller, J. Bresnan and M. Halle (eds.), *Linguistic Theory and Psychological Reality.* Cambridge, Mass., MIT Press.

McMaster, I., J. Sampson and J. King (1976) Computer acquisition of natural language: A review and prospectus. *Intern. J. Man-Machine Studies, 8,* 367–396.

McNeill, D. (1966) Developmental psycholinguistics. In F. Smith and G. Miller (eds.), *The genesis of language.* Cambridge, Mass., MIT Press.

Miller, G. (1967) Project Grammarama. In *The Psychology of Communication.* Hammonsworth, NY: Basic Books.

Moeser, S. and A. Bregman (1972) The role of reference in the acquisition of a miniature artificial language. *J. verb. Learn. verb. Behav., 12,* 91–98.

Moeser, S. and A. Bregman (1973) Imagery and language acquisition. *J. verb. Learn. verb. Behav., 12,* 91–98.

Newell, A. and H. Simon (1973) *Human problem solving.* Englewood Cliffs, N.J., Prentice Hall.

Newport, E., H. Gleitman and L. Gleitman (1977) Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow and C. Ferguson (eds.), *Talking to Children: Input and Acquisition.* New York, Cambridge University Press.

Norman, D. and D. Rumelhart (1975) *Explorations in Cognition.* San Francisco, W. H. Freeman.

Peters, S. and R. Ritchie (1973) On the generative power of transformational grammars. *Infor. Sci., 6,* 49–83.

Postal, P. (1964) Limitations of phrase structure grammars. In J. Fodor and J. Katz (eds.), *The Structure of Language.* Englewood Cliffs, N.J., Prentice Hall.

Putnam, H. (1971) The "Innateness Hypothesis" and explanatory models in linguistics. In J. Searle (ed.), *The Philosophy of Language.* London, Oxford University Press.

Reeker, L. (1976) The computational study of language acquisition. In M. Yovits and M. Rubinoff (eds.), *Advances in Computers, Vol. 15.* New York, Academic Press.

Rochester, S. (1973) The significance of pauses in spontaneous speech. *J. Psycholing. Res., 2,* 51–81.

Schlesinger, I. (1971) Production of utterances and language acquisition. In D. Slobin (ed.), *The Ontogenesis of Grammar.* New York, Academic Press.

Siklóssy, L. (1971) A language learning heuristic program. *Cog. Psychol., 2,* 279–295.

Siklóssy, L. (1972) Natural language learning by computer. In H. Simon and L. Siklóssy (eds.), *Representation and Meaning: Experiments with Information-processing Systems.* Englewood Cliffs, N.J., Prentice Hall.

Sinclair de-Zwart, H. (1969) Developmental psycholinguistics. In D. Elkind and J. Flavell (eds.), *Studies in Cognitive Development: Essays in Honor of Jean Piaget.* New York, Oxford University Press.

Slobin, D. (1973) Cognitive prerequisites for the development of grammar. In C. Ferguson and D. Slobin (eds.), *Studies in Child Language Development.* New York, Holt, Rinehart and Winston.

Slobin, D. (1978) Universal and particular in the acquisition of language. In *Language Acquisition: State of the Art.* Conference at the University of Pennsylvania, Philadelphia, May 1978.

Snow, C. (1972) Mothers' speech to children learning language. *Child Devel., 43,* 549–565.

Snow, C. and C. Ferguson (1977) *Talking to children: Language Input and Acquisition.* New York: Cambridge University Press.

Solomonoff, R. (1964) A formal theory of inductive inference. *Infor. Control, 7,* 1–22; 224–254.

Soloway, E. and E. Riseman (1977) Levels of pattern description in learning. (COINS Technical Report 77-5), Computer and Information Science Dept., Amherst, Mass., University of Massachusett.

Van der Mude, A. and A. Walker (1978) On the inference of stochastic regular grammars. *Infor. Control, 38,* 310–329.

Wexler, K., P. Culicover and H. Hamburger (1975) Learning-theoretic foundations of linguistic universals. *Theoret. Ling., 2,* 215–253.

Wexler, K. and H. Hamburger (1973) On the insufficiency of surface data for the learning of transformational languages. In K. Hintikka, J. Moravcsik and P. Suppes (eds.), *Approaches to Natural Languages.* Dordrecht, Netherlands: Reidel.

Wharton, R. (1974) Approximate language identification. *Infor. Control, 26,* 236–255.

Wharton, R. (1977) Grammar enumeration and inference. *Infor. Control, 33,* 253–272.

Winograd, T. (1972) A program for understanding natural languages. *Cog. Psychol., 3,* 1–191.

Winston, P. (1975) Learning structural descriptions from examples. In P. Winston (ed.), *The Psychology of Computer Vision.* New York, McGraw-Hill.

Woods, W. (1975) What's in a link: Foundations of semantic networks. In D. Bobrow and A. Collins (eds.), *Representation and Understanding: Studies in Cognitive Science.* New York, Academic Press.

Résumé

Analyse d'une recherche centrée sur l'apprentissage du langage humain, développant des modèles mécanistes précis susceptibles, en principe, d'acquérir le langage à partir d'une exposition aux données linguistiques. Une telle recherche comporte des théorèmes (empruntés à la linguistique mathématique) des modèles informatiques pour l'acquisition du langage (empruntés à la simulation cognitive et à l'intelligence artificielle) des modèles d'acquisition de la grammaire transformationnelle (empruntés à la linguistique théorique). On soutient que cette recherche repose étroitement sur les thèmes principaux de la psycholinguistique de développement et en particulier sur l'opposition nativisme-empirisme, sur le rôle des facteurs sémantiques et pragmatiques dans l'apprentissage du langage, sur le développement cognitif et l'importance du discours simplifié que les parents adressent aux enfants.