CMU-CS-81-129

A Theory of Language Acquisition Based on General Learning Principles

John R. Anderson
Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

This project is supported by contract N00014-79-C-0661 from the Office of Naval Research. I would like to thank Steven Pinker, Lynne Reder, and Miriam Schustack for their comments on earlier drafts of this manuscript.

Abstract

A simulation model is described for the acquisition of the control of syntax in language generation. This model makes use of general learning principles and general principles of cognition. Language generation is modelled as a problem solving process involving principly the decomposition of a to-be-communicated semantic structure into a hierarchy of subunits for generation. The syntax of the language controls this decomposition. It is shown how a sentence and semantic structure can be compared to infer the decomposition that led to the sentence. The learning processes involve generalizing rules to classes of words, learning by discrimination the various contextual constraints on a rule application, and a strength process which monitors a rule's history of success and failure. This system is shown to apply to the learning of noun declensions in Latin, relative clause constructions in French, and verb auxiliary structures in English.

This paper reports the current state of a theory about the acquisition of the syntax in natural language generation. This theory is intended to appy to inductive learning (learning from examples) by either adults or children.

A serious question exists in computational linguistics as to whether it is necessary to deal with the full complexity of syntax in order to comprehend language (e.g., Schank, 1975; Birnbaum & Selfridge, 1979). Conceptual and knowledge-based approaches to language parsing often seem much more efficient. However, it seems hard to deny that a language generation system must have full grasp of the syntax of language and it is hard to deny that relatively young children are successful at obtaining a grasp of this syntax. Consistent with this view that syntax is more important to generation than to comprehension is some of the recent evidence that children appear to display more intricate knowledge of syntax in generative tests than receptive tests (Schustack, 1979). Therefore, in this research I have focused on the acquisition of generative capacity. However, I think the same learning mechanisms would apply to acquisition of a receptive capacity, but I think the receptive system so acquired would rely less on syntax than the generative system.

This research has its background in past work on language acquisition (for reviews, see Anderson, 1976; Pinker, 1979--see also Langley, 1981), especially in my previous work on LAS (Language Acquisition System--see Anderson, 1977). For various reasons that will be explained, there were problems with LAS and a more general concept of human cognition was developed called ACT (Anderson, 1976). The system to be reported here is an attempt to merge the ideas in the ACT project and the LAS project. It is called ALAS for ACT's Language Acquisition System. First in this paper I will review those aspects of the LAS and ACT systems that are relevant to understanding the current project and then I will turn to describing the ALAS system.

The LAS System

LAS accepted as input strings of words, which it treated as sentences, and scene descriptions encoded as associative networks. When learning, the program attempted to construct and modify augmented transition networks which described the mapping between sentence and scene descriptions. This assumption, that the program has access to sentence-meaning pairings, is the basic assumption underlying most of the recent attempts at language acquisition. This assumption might be satisfied in the circumstance where the child is hearing a sentence describing a situation he is attending to. Even here it is likely that the child will represent aspects of the situation not described and fail to represent aspects described. In LAS we worked out mechanisms for filtering out the non-described aspects of the meaning representation by comparison with the sentence. In the current ALAS system there is a discrimination mechanism for bringing in aspects of the situation not initially thought by the learner to be part of the sentence. So, we have worked out mechanisms for achieving sentence-meaning pairings in simple ostensive learning situations. However, much of what a child must learn about language will lack simple ostensive referents. For instance, most of the verb auxiliary system refers to non-ostensive meaning. How a child (or any system) would come up with sentence-meaning pairings in these situations is not clear and remains an issue for future research.

A major assumption of the LAS model that is maintained in the current system is that the system already knows the meaning of a base set of words. LAS was unable to learn the meaning of any

words in context while the current system can; however the basic learning algorithm in both still requires that a substantial number of words in the sentence have their meanings previously learned. In principle (see Anderson, 1974), it would be possible to call to bear statistical learning programs to extract the meaning of the base set of words from a sufficiently large sample of meaning-sentence pairings. However, the evidence (McWhinney, 1980) is that children accomplish their initial lexicalization by having individual words paired directly with their referents.

Identifying Phrase Structure: The Graph Deformation Condition

A major problem in language learning is to identify the phrase structure of the sentence. There are a number of reasons why inducing the syntax of language becomes easier once the phrase structure has been identified: (1) Much of syntax is concerned with placing phrase units within other phrase units. (2) Much of the creative capacity for generating natural-language sentences depends on recursion through phrase structure units. (3) Syntactic contingencies that have to be inferred are often localized to phrase units, bounding the size of the induction problem by the size of the phrase unit. (4) Natural language transformations are best characterized with respect to phrase units as the transformational school has argued. (5) Finally, many of the syntactic contingencies are defined by phrase unit arrangements. So, for instance, the verb is inflected to reflect the number of the surface structure subject.

A major mechanism for identifying phrase structure in LAS (and which is continued in ALAS) is use of the graph-deformation condition. The idea is to use the structure of a sentence's semantic referent to place constraints on surface structure. The application of the graph deformation condition is illustrated in Figure 1. In part (a) we have a semantic network representation for a series of propositions and in part (b) we have a sentence that communicates this information. The network structure in (a) has been deformed in (b) so that it sits above the sentence but all the node-to-node linkages have been preserved. As can be seen, this captures part of the sentence's surface structure. At the top level we have the subject clause (node X in the graph), *gave*, *book*, and the recipient (node Y) identified as a unit. The two noun phrases are segmented into phrases according to the graph structure. For instance, the graph structure identifies that the words *lives* and *house* belong together in a phrase and that *big*, *girl*, *lives*, and *house* belong together in a higher phrase.

The graph deformation in part (b) identifies the location of the terms for which meanings are possessed in the surface structure of the sentence. However, terms like *the* before *big girl* remain ambiguous in their placement. It could either be part of the noun phrase or directly part of the main clause. Thus, there remains some ambiguity about surface structure that will have to be resolved on another basis. In LAS the remaining morphemes were inserted by a set of ad hoc heuristics that worked in some cases and not in others. One of the goals in ALAS was to come up with a better set of principles for determining the boundaries of phrases.

The graph deformation condition is violated by certain sentences which have undergone structure-modifying transformations that create discontinuous elements. Examples in English are:

1. The news surprised Fred that Mary was pregnant.

2. John and Bill borrowed and returned, respectively, the lawnmower.

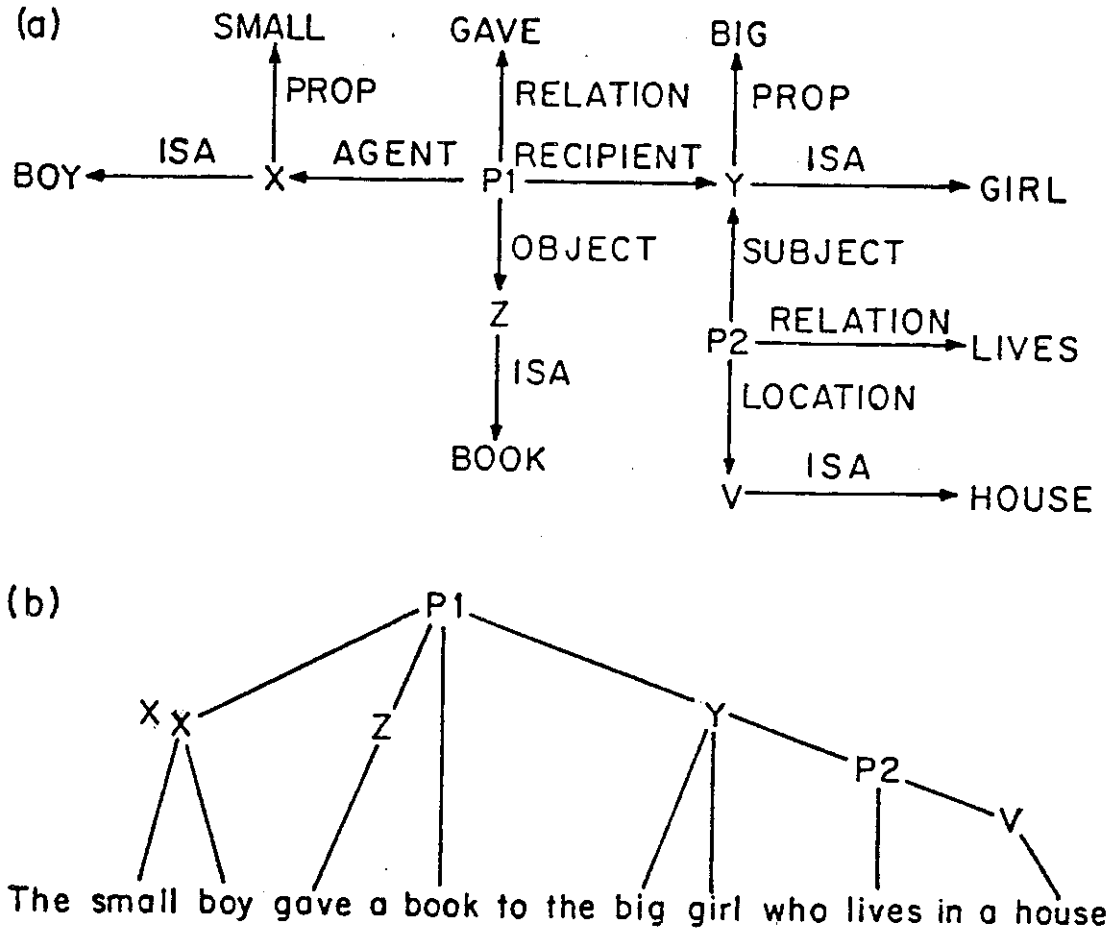

Figure 1

Transformations which create discontinuous elements are more common in languages that use word order less than English. However, the graph deformation condition remains as a correct characterization of the major tendency in all languages. The general phenomena has been frequently commented upon and has been called Behaghel's First Law (see Clark & Clark, 1977). A problem with LAS was that it had no means of dealing with exceptions to the graph deformation conditions or of learning transformations in general. Another goal for the ALAS current enterprise is to be able to detect sentences that violate the graph deformation condition and to use these as opportunities for learning transformations.

A major source of my dissatisfaction with LAS is that its processing discipline and learning mechanisms are specific to language and it was hard to imagine how they would relate to other types of skill learning. While many people believe the principles underlying language acquisition are unique, I do not. I think the other problems with the LAS enterprise could be repaired but I felt a fresh start was needed if we were to show that general skill acquisition principles could plausibly apply to natural language as a special case. This led to the development of the ACT theory (Anderson, 1976; Anderson, Kline, & Lewis, 1977) and to a set of learning principles for that theory.

ACT

As originally formulated, ACT was a production system without any commitment to the mechanisms of skill organization or skill acquisition. However, a set of principles have emerged in our more recent work (Anderson, Kline, & Beasley, 1980; Anderson & Kline, 1979; Anderson, Greeno, Kline, & Neves, 1981) and it is these developments which are essential for the current application. These ideas have been developed in non-linguistic domains--schema abstraction, acquisition of proof skills in geometry, and most recently in the acquisition of programming skills.

We see any skill as being hierarchically organized into a search of a problem space in which there is a main goal, which is decomposed into subgoals, and so on until the decomposition reaches achievable subgoals. Much of what is distinctive about a particular skill is the way in which the problem space is searched for a solution. In our model of language generation, this is seen as a simple top-down generation of subgoals (corresponding to phrases) where there is no real search needed unless transformations have to be applied. We will illustrate this application to language shortly.

In simulating language acquisition we have focused on the learning mechanisms concerned with operator selection: generalization, discrimination, and strengthening. Generalization takes rules developed from special cases and tries to formulate more general variants. Discrimination is responsible for acquiring various contextual constraints to delimit the range of overly general rules. Strength reflects the success of a rule in the past and controls its probability of future application. In combination, these mechanisms function like a statistical learning procedure to determine which problem features are predictive of a rule's success. They have been extensively documented in our efforts to model the literature on schema abstraction (Anderson & Kline, 1979; Elio & Anderson, in revision), but they have had a richer application to acquisition of proof skills (Anderson, submitted; Anderson, Greeno, Kline, & Neves, 1981). I will sketch their application to the language acquisition domain, but the reader should go to these other sources (and particularly Anderson, Kline, & Beasley, 1980) for a fuller development.

Current Framework for Language Learning

The language learner is characterized as having the goal of communicating a particular set of propositions. This set of propositions is organized into a main proposition and subpropositions. So, for instance, the goal behind the generation of *The girl kicks the boys* might be a communication structure which we can represent as (KICK (GIRL x) (BOY y)) where x is tagged as singular and y is tagged as plural. To achieve the goal, the learner tries to decompose this higher level goal into subgoals, according to the units of the overall communication structure. So, he will decompose this into the subgoals of communicating *kick*, of communicating (GIRL x), and of communicating (BOY y). He looks to his language for some means of organizing these subgoals. So, he might have learned a rule of the form:

IF the goal is to communicate (LVrelation LVobject1 LVobject2)
and LVrelation is in the VERBX class

THEN set as the subgoals to
communicate LVobject1
say the morpheme for LVrelation
say "s"
and to communicate LVobject2

or we might more compactly denote this rule:

(1 2 3) --> 2 + 1* + S + 3 if 1 in VERBX

In the above, the 1, 2, and 3 match the three elements in the meaning structure--KICK, (GIRL x), and (BOY y). The right side of the arrow specifies their order in the sentence and the insertion of mophemes like *S*. The star above the 1 indicates its lexical form is to be retrieved. The other elements will have to be further unpacked.

If it is early in the language learning history and the learner does not have a rule for realizing this construction, then he might try to invent some principle. He may only produce a fragment (e.g., girl hit) or a non-allowed order (e.g., girl boy hit). There is some evidence in first language acquisition that children will use word orders not frequent in adult speech (Clark, 1975; de Villiers & de Villiers, 1978; McWhinney, 1980). For instance, there is a tendency to prefer agents first even when one's language does not. Also, it is well known that second language learners fall back on their first language word orders when knowledge of word order fails.

The embedded subgoals are unpacked into actions or further subgoals in the same way that the top level structure is unpacked. For instance, if the object to be communicated were (girl x (like x (sailor z))), the top level of this structure might be communicated by the rule:

(1 2 3) --> the + 1* + 3 if 1 is a noun

where (like x (sailor z)) is item 3 in the above and would be communicated by the rule:

(1 2 3) --> who + 1* + 3 if 1 is a verb
and the construction is embedded

Figure 2 illustrates the hierarchy of subgoals in the generation of a relatively complex sentence: *The young policeman sees the lawyer whom the crook paid*. It should be clear that if sentences are generated by setting subgoals to reflect the structure of the referent, then the graph deformation condition will tend to be satisfied in natural language.

A set of interesting questions arise when we try to augment this system with a set of performance assumptions about how many subgoals the system can maintain in working memory and whether it has rules readily available for decomposing the goals or has to try to invent rules in generation. In these performance assumptions would lie an account of the telegraphic speech of young children (in which much information is omitted from sentences--see de Villiers & de Villiers, 1978). However, the work to be reported has ignored the existence of possible performance limitations and has assumed an ability to sustain arbitrarily complex structures. This simplification allows us to focus on the general competence of the learning system.

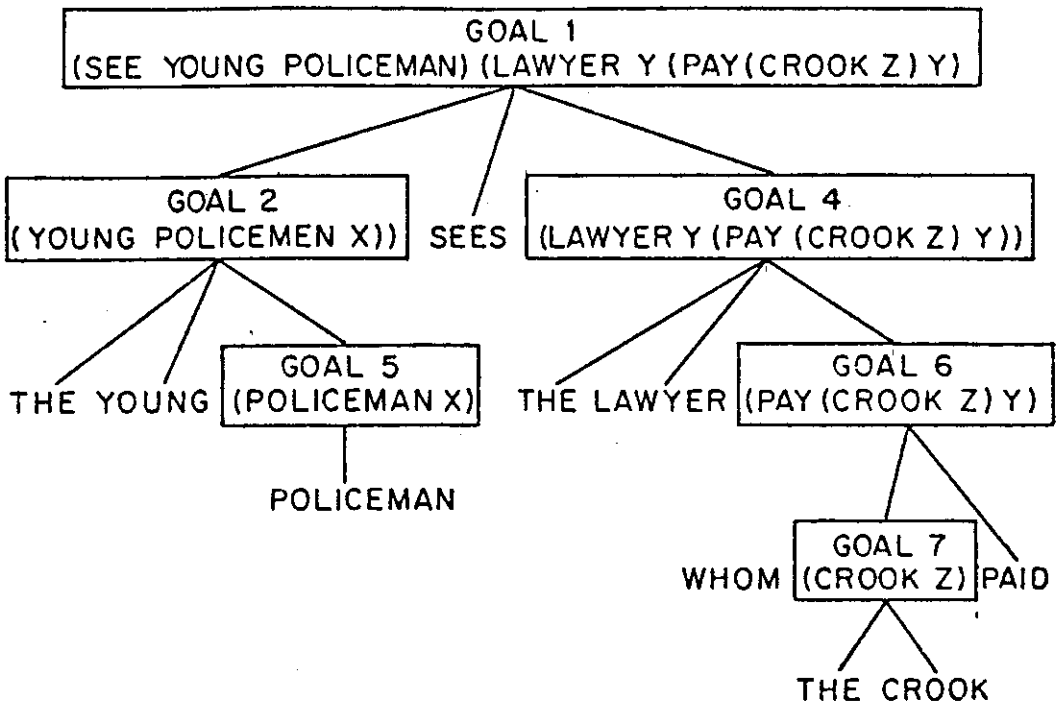

Figure 2

The learning that occurs in ALAS is basically learning by doing. The learner generates an utterance and it is assumed that he has access to feedback about the correctness of the construction he generated and perhaps information about what the correct utterance should have been if he has made an error. There are many ways this can happen. The learner may generate a sentence and be corrected by a teacher. He may generate a sentence and remember a sentence or sentence fragment heard earlier. He may hear a sentence, infer its meaning, and compute how he would express the meaning. By whatever means the learner sometimes identifies some fragment of his generation to be in error and sometimes has a hypothesis as to the correct utterance. This is the stimulus for learning. In the actual simulations that will be reported, the program is given a model sentence along with each meaning and the program compares its generation with the model sentence. No doubt this is an unrealistically ideal assumption and results in a considerable speed up of the learning process in ALAS. However, the same learning mechanisms would apply in more psychologically realistic situations where the program was given only occasional information and often fragmentary information about what the correct target sentences were.

Formation of Initial Rules

The initial rules that the system acquires are, of course, quite specific. So, for instance, consider the rules it might form upon receiving a pairing of the Latin sentence ((Equ i)(agricol as)port ant) and the meaning representation (carry (horse x)(farmer y)). With a partially complete lexicalization, ALAS knows the meaning of *equ* is horse, the meaning of *agricol* is farmer, and the meaning of *port* is carry). ALAS then formulates the following rules:

```
(1  2  3) -->  2 + 3 + 1* + ant        if 1 = carry
(1  2)    -->  1* + i                  if 1 = horse
(1  2)    -->  1* + as                 if 1 = farmer
```

Thus, its acquired rules are exact encodings of the relations at each level in the meaning hierarchy. The evidence is that children also start out with rules specific to individual words (MacWhinney, 1980; Maratsos & Chalkley, 1981) and indeed the nature of natural language makes this a wise policy in that rules are quite specific to various lexical items (Bresnan, 1981; Maratsos & Chalkley, 1980; Pinker, 1981). This also is exactly how learning proceeds in other areas to which we have applied ACT. Initially, the system acquires rules that encode the exact goal structure of specific examples. Later, generalizations are formed.

While, on one hand, these rules are too specific, on the other hand, they are too general. The inflections associated with the nouns and verbs are only correct for the specific case and number combinations but these rules do not reflect that constraint. The system will have to acquire discriminating features that will properly constrain the range of application of these rules. Again that corresponds to child language. Children initially use words with a single inflection in all situations and only later acquire the contextual constraints. It also corresponds to our other learning endeavours where contextual constraints on goal decomposition are acquired through discrimination.

Discrimination

To illustrate the discrimination process consider again the rule for realizing *farmer*:

```
(1  2)  -->  1* + as        if 1 = farmer        (a)
```

Suppose the system encounters a second instance of farmer in the meaning-sentence pairing (call (farmer u) (girl v)) · ((agricol a) (puell am) voc at). It would detect a conflict between its generation of *agricol + as* and the target *agricol + a*. In this case it would look for differences between the context of its current application and the previous. The relevant differences are:

1. *y* in the previous application is tagged as plural while *u* in the above structure is singular.

2. The object structure was in third position in the embedded clause of the first meaning structure, but now it is in second position.

However, there are any number of other potential differences such as

3. The previous verb was *port* and the current *voc*.

4. The second position of the embedded clause was plural and the current is singular.

5. The current sentence involves a feminine object.

LAS has an ordering of distance (to-be-explained) such that 4 and 5 above would be definitely less preferred but there is no clear basis for choosing 1 and 2 over 3. A feature to discriminate upon is chosen at random and a new rule is formed such as:

```
(1  2)  -->  1* + a        if 1 = farmer                          (b)
                              and 2 is singular
```

Note that this is a discrimination for the current context, not the previous. ALAS can also form a rule for the old context

```
(1  2)  -->  1* + as       if 1 = farmer                          (c)
                              and 2 is plural
```

but only if the old rule (a) exceeds a threshold of strength to indicate that it has applied successfully more often than not and is therefore not a pure mistake.[1]

The correct rules above need another round of discrimination before they pick up the semantic position feature. Then they will become

```
(1  2)  -->  1* + a    if 1 = farmer                              (d)
                          and 2 is singular and this occurs in second
                          position in the semantic referent

(1  2)  -->  1* + as   if 1 = farmer                              (e)
                          and 2 is plural and this occurs in third
                          position in the semantic referent
```

The set of possible features for discrimination is defined by a network that includes the semantic referent, the goal structure, and any properties tagged to terms in the semantic referent or the goal structure. The program does a breadth first search out from the current position in this network looking for features that distinguish between current and past applications of the rule. It chooses the features it first finds in that search. This means that the system is sensitive to both syntactic and semantic contingencies of the context of application.

Generalization

Let us consider the production form of the rule (e) from above:

> IF the goal is to communicate LVobject2 = (farmer LVterm)
> and LVterm is plural

---

[1] If the discrimination process chooses an incorrect feature as in

```
(1  2) --> 1* + a    if 1 is farmer
                        and the structure is in the context of port
```

this rule will not lead to worse performance than the original and will eventually lose out as the correct discriminations are formed.

and the higher goal is (LVrelation LVobject1 LVobject2)
THEN generate agricol + as

Another production would be:

IF the goal is to communicate LVobject2 = (girl LVterm)
and LVterm is plural
and the higher goal is (LVrelation LVobject1 LVobject2)
THEN generate puell + as

An application of the generalization mechanism in ACT would yield:

IF the goal is to communicate LVobject2 = (LVclass LVterm1)
and LVword is the word for LVclass
and LVterm is plural
and the higher goal is (LVrelation LVobject1 LVobject2)
THEN generate LVword + as

or in our compressed notation

```
(1 2)  -->  1* + as        if 2 is plural
                              and this occurs in third position of the
                              semantic referent
```

where we have changed the restriction that it apply to a particular word to allow anything that fits a pair of variables (LVclass, LVword). This would lead to an enormous overgeneralization in that the above rule is only valid for first-declension nouns.

Of course, we do not know how Latin was acquired, but the evidence for other languages (Maratsos & Chalkley, 1981) is against the existence of such rampant overgeneralizations. Some overgeneralizations do occur (and they can in ALAS) but what is remarkable is their lack of frequency. Certainly, overgeneralizations are much less frequent than would be produced by the above mechanism. What is more common is undergeneralization where children first generalize a rule to a much smaller range of terms than that to which it can apply.

Thus, we have had to assume that generalization cannot occur in language by the wholesale replacement of a constant by a variable. Rather what we assume is that generalization occurs by replacing a constant by a word class. So, the proper form of the above rule becomes

```
(1 2)  -->  1* + as        if 1 is in class X
                              and 2 is plural and this occurs in third
                              position in the semantic referent
```

where class X will contain *farmer* and *girl* among others. It is unclear at present whether this is a true instance of where language acquisition differs from other cognitive learning or whether the generalization mechanism should be set up to produce constrained variables in all situations.

A major issue in ALAS concerns when words should be merged into the same class. It is not the

case that this occurs whenever there is the potential to merge two rules as above. The existence of overlapping declensions and overlapping conjugations in many languages would result in disastrous overgeneralizations. Rather we have brought to bear an extension of our schema abstraction ideas (Anderson & Kline, 1979). What ALAS does is look at the pattern of rules that individual words appear in. It will merge two words into a single class when

1. The total strength of the rules for both words exceeds a threshold indicating a satisifactory amount of experience

2. A fraction (currently 2/3) of the rules that have been formed for one word (as measured by strength) have been formed for the other word.

When such a class is formed, the rules for the individual words can be generalized to that class. Also, any new rules acquired for one word will generalize to the other. Once a class is formed new words can be merged with the class according to the same criteria (1) and (2) for merging words. Further, two classes can be merged together, again according to the same criteria. Thus, it is possible to gradually build up large classes like first declension.

The word-specific rules are not lost when the class generalizations appear. Furthermore, one form of discrimination is to propose that there is a rule special to a word. Because of the specificity ordering in production selection, these word-specific rules will be favored when applicable. This means that the system can live with a situation where a particular word (such as *dive*) can be in a general class but still maintain some exceptional behavior.

Thus, the system begins with a lot of word-specific rules which gradually expand in their scope of application. This is basically the development observed in child language.

It should be noted that there is another dimension in which the system's behavior starts out very general. The rules for communicating a particular construction, such as an object construction (e.g. noun phrase) or qualifying proposition (e.g., a relative clause), are assumed to apply in every location. Thus, the system automatically assumes rules are recursive and does not need, as did LAS, to verify such points of recursion. Rather, the learning here takes the form of constraining this assumption where overgeneral--as we have discussed. Correspondingly, children seem not reluctant to venture old constructions in new syntactic contexts.

Phrase Structure Segmentation

Up to this point we have assumed that the target sentences were segmented into phrase structure units. The graph deformation condition can be used to assign the words whose meaning is known to phrase units but this leaves unspecified the other morphemes. To take an example from my work with Latin consider the following meaning-sentence pairing:

(praise (friend u (have (man v) u)) (field x (have (farmer y) x))) 1
amic us vir i ager os agricol ae laud at 2
(translated: The man's friend praises the farmer's fields).

Clearly, the semantic structure indicates *vir* (man) associates with *amic* (friend) as a modifier and not

with *ager* (field) since *man* is contained in the same meaning unit as *friend*. However, the semantic structure provides us no way of deciding whether the non-meaning-bearing morpheme *us* associates with *vir* or *amic*. Similarly, it is ambiguous how to locate the other noun inflections: *i*, *os*, and *ae*. On the other hand, *at* occurring at the end of the sentence definitely must associate with *laud*. Thus, by means of the graph deformation condition and only taking unambiguous cases, we get the following hierarchical organization for the Latin string:

((amic ((vir))) (ager ((agricol))) laud at) 3

where the indeterminate morphemes are left out. At one point in its application of the graph deformation condition ALAS calculates just this structure. If nothing more can be done, this is the form of the string provided to the learning system--i.e., with the ambiguous morphemes deleted.

How can this string be improved upon to insert the non-meaning bearing morphemes? In the literature there are three suggestions. First, there may be pauses in the speech signal to indicate the correct associations. There would be no ambiguity if there were long pauses after *us*, *i*, *os*, and *ae* in the above message. Normal speech does not always have such pauses in correct places and sometimes has pauses in wrong places. Still, this basis for segmentation would be correct more often than not and ALAS's error correcting facilities have the potential to recover from the occasional missegmentaion. Also, it is argued that parent speech to children is much better segmented than adult speech to adults (see de Villiers & de Villiers, 1978). In ALAS pausing is used when given, but the system does not require pause segmentation.

A second suggestion is to use past instances of successful segmentation to segment in the current case. Thus, if the system has previously identified *agricol + ae* as associating together it can assume they associate together now. The past experience could derive from hearing the word in isolation or from other sentences where some other basis could be applied for segmentation. Memory for words spoken in isolation is a particularly useful solution to the problem of identifying which morphemes belong together to define a word. The evidence is quite clear that children do hear many words in isolation (McWhinney, 1980). This is less helpful in identifying phrase boundaries for structures like noun phrases or relative clauses--both because these structures are less likely to be spoken in isolation and because the same word sequence is rarely repeated. This may explain why missegmentation of morphemes within words is rare in child speech relative to missegmentation of words with phrases (Slobin, 1973). Although we could in principle use this strategy, our simulation that attempted to segment without pause structure was not given words in isolation.

The third basis for segmentation relies on the use of statistics about morpheme-to-morpheme transitions. For instance, the segment *ae* will more frequently follow *agricol* with which it is associated than it will precede *laud* with which it is not. The differences in transitional frequencies would be very sharp in a language like Latin with a very free word order but they also exist in English. Thus, ALAS can associate *ae* with the *agricol* if it has followed *agricol* more frequently than it has preceded *laud*. This requires keeping statistics about word-to-word transitions. Currently, the system will favor one association of a morpheme over another if there is a difference in frequency of two. This might seem a rather small threshhold but I have gotten satisfactory performance out of ALAS, partly because ALAS can recover from occasional missegmentations. Again the evidence is that children do occasionally missegment (McWhinney, 1980) and, of course, recover eventually. It strikes some as implausible to suppose that people could keep the statistical information required about

word to word transitions. However, Hayes and Clark (1970) have shown that subjects in listening to nonsense sound streams can use differential transition probabilities as a basis for segmentation. Such information has also proven useful in computational models of speech recognition (Lesser, Hayes-Roth, Birnbaum, & Cronk, 1977).

It is possible and frequently has been the case that none of the ALAS segmentation mechanisms could apply to assign a morpheme to a level in the phrase structure. In such cases the non-assigned morpheme was simply omitted from the phrase structure. Thus, the initial utterances produced by ALAS, like the utterances produced by young children, are telegraphic in character. That is, they are missing many functors.

Having now described the basic learning principles embedded in ALAS, I would like to describe their application in three simulation efforts. Each focused on a different aspect of language and each illustrates different features about ALAS.

Latin: The issue of segmentation

Our first endeavour was to learn a fragment of Latin that involved first and second declension nouns, inflected for the nominative, accusative, and genitive cases and for plural and singular. An example of the input to ALAS is

> Agricol ae puel am legat i laud ant
> (praise (farmer x) (girl u (have (lieutenant v) u)))
> where x is plural, u and v are singular

That is, the input was a string of Latin morphemes that comprised the target sentence and a hierarchical representation of the meaning of this sentence. The program was provided with a long sequence of such pairings. Over the sequence all syntactic possibilities were realized. With each pairing, ALAS consulted its rules to see if they would map the meaning structure onto the target string. Its learning principles were evoked to modify the rules if they failed to produce the right mapping. As can be seen, in this simulation (and the others) we provide the strings segmented into morphemes. Acquisition of morpheme segmentation is thus being ignored. The verbs used were *8* first-conjugation verbs; the nouns were *8* first-declension nouns and *7* second-declension nouns. One of the things our simulation was going to get at was the adequacy of our class heuristics to separate our first and second declension nouns. We performed two simulations over this target language subset. In the first we provided the system with no information about segmentation and it was forced to use the graph-deformation condition and transitional probabilities to segment into surface structure units. In the second simulation we provided pause information to indicate with which words the inflections were associated.

To avoid any possible biasing in input order, the sentence-meaning pairs were generated by a randomization program. The simulation without the pause information required 525 pairings before it has identified all the needed grammatical rules and ran a criterion 25 pairings with no mispredictions of the target strings. With pause information, only 100 sentences were required to reach the same criterion. Figure 3 illustrates the mean number of errors for the two conditions plotted as a function of the logarithm of number of pairings experiences. An error was defined as a misordering of elements

at any phrase level, the insertion of an incorrect morpheme, or the ommission of a morpheme.

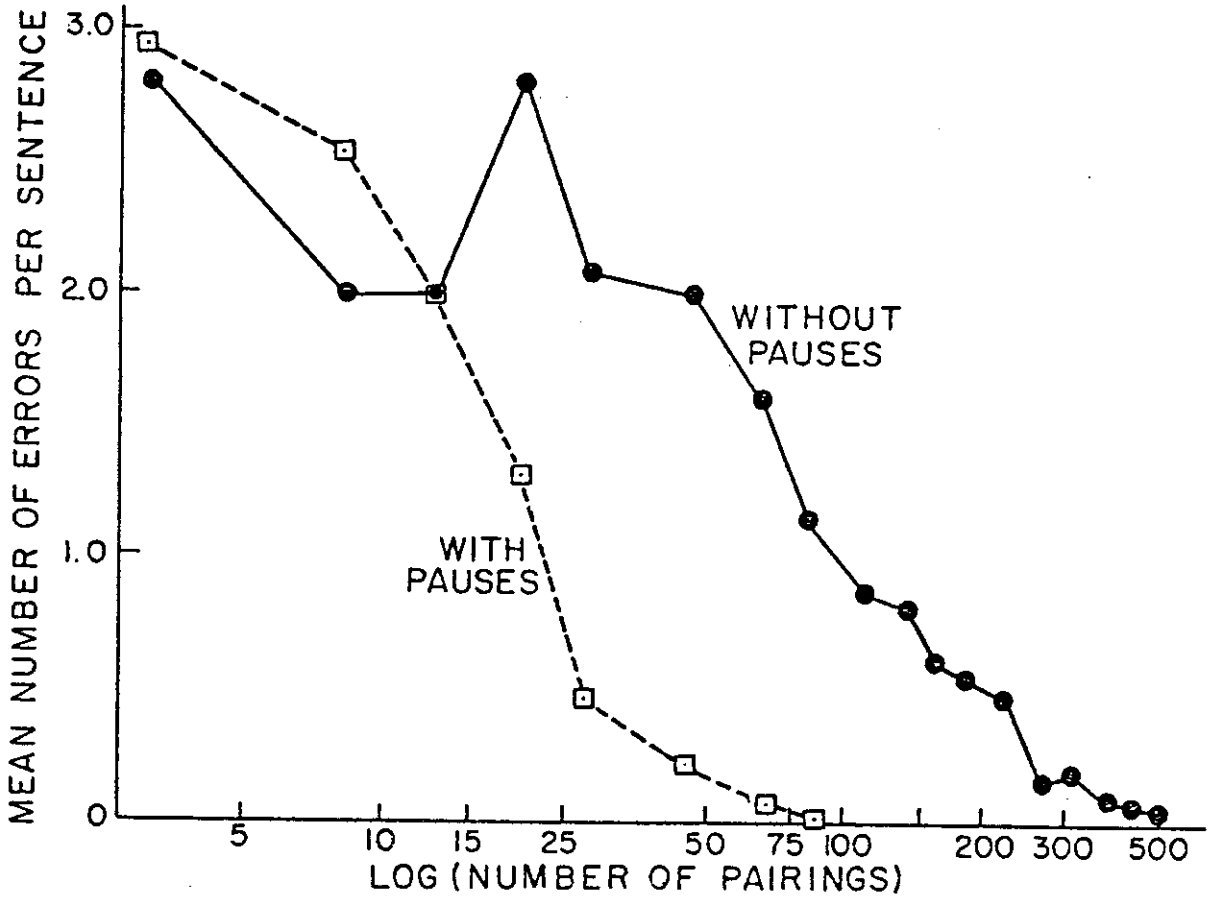

Figure 3

In the case where the system was not given information about pause structure, it had to use transitional frequency to segment. After the first 25 sentences it was correctly associating about 50% of the noun inflections with the nouns. Most of the remaining 50% were failures to insert the morphemes but there were occasional missegmentations. Despite the fact that it was correctly segmenting over half of the input to the learning program after the 25th trial, it was only after 75 trials that any learning of inflections showed up in its performance (i.e., it started using these inflections with significantly greater than chance accuracy). Even after 150 sentences ALAS is failing to segment some nouns in 10% of the sentences. The difficulty in segmentation is what is accounting for the slow learning of the program. The examples that follow present, first, the Latin morpheme string that the program generated to express a meaning structure (not shown) and, second, the target string that was correct. I have given a non-random selection of these to give the reader a sense of the progress of the system throughout the course of the 525 pairings:

Sentence 2: PUGN NUNTI LEGAT
vs. NUNTI I LEGAT OS PUGN ANT
Sentence 17: NUNTI TUB LAUD ANT
vs. NUNTI I TUB UM LAUD ANT

Sentence 28: AGRICOL PUELL AE LEGAT AM ANT
vs. AGRICOL AE PUELL AM LEGAT I AM ANT
Sentence 52: FEMIN VIR OS LAUD AT
vs. FEMIN A VIR OS LAUD AT
Sentence 83: LEGAT I POET A NUNTI I LAUD ANT
vs. LEGAT I POET AM NUNTI ORUM LAUD ANT
Sentence 129: LEGAT US NUNTI UM AM SPECT AT
vs. LEGAT US NUNTI UM SPECT AT
Sentence 203: VIR I AMIC OS NATUR AE OCCUP ANT
vs. VIR I AMIC OS NATUR AE OCCUP ANT
Sentence 429: AMICIT AS AGRICOL AS PUGN ANT
vs. AMICIT AE AGRICOL AS PUGN ANT

The class formation heuristics worked quite well in these simulations. Both with and without pause information, the two declensions were identified as two word classes and all the verbs were brought together into another word class. Figure 4 illustrates the history of discrimination that led to correct use of inflections for the second declension in the simulation with pause information. Time goes to the right and down in the figure. It turned out that on four occasions the system proposed an unconstrained rule for the *us* inflection. This is reflected in the horizontal dimension. Going down we have the history of discrimination for each rule. Arrows lead from a rule to a discriminated rule. The label on the arrow indicates the feature added in the discrimination. Thus, for instance, A35 is a rule that calls for the *us* inflection (appropriate for nominative singular). It was used incorrectly in an accusative plural situation and an *os* rule, A66, was formed with the discriminating test that the noun be in accusative case (i.e., third position in the semantic structure). This rule misapplied in an accusative singular situation and so a singular feature was added. Rules in boxes are ones that were so weakened by misapplication that they were removed.

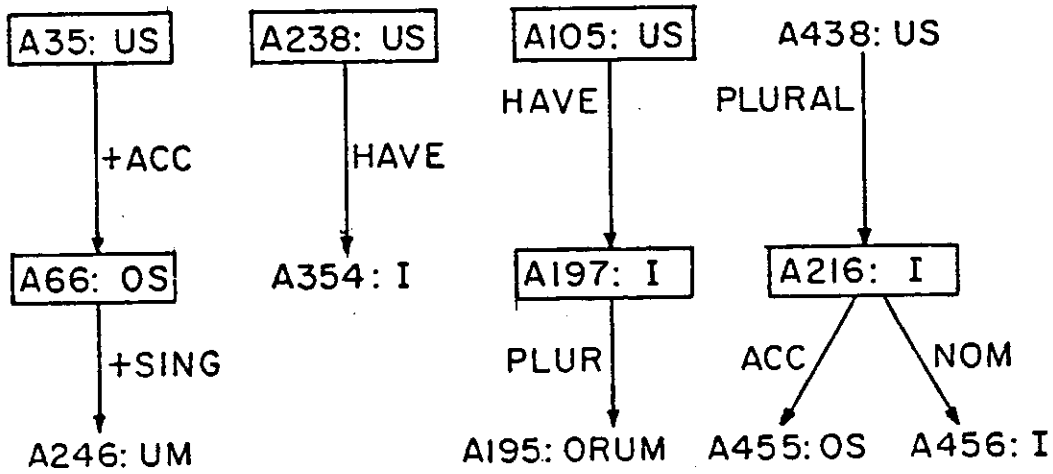

Figure 4

Note that there are four rules with all the necessary features: A246 for accusative singular, A195 for genitive plural, A455 for accusative plural, and A456 for nominative plural. On the other hand, A354 for genitive singular only tests that it is in a possessive context and not for number. However, because of the specificity ordering on production selection, the more specific genitive plural rule

(A195) will apply whenever applicable leaving A354 only the genitive singular situations in which to apply. Similarly, A438 which has no discriminating features will only apply when no other rule is applicable--which is to say it will apply only the nominative singular case for which it gives the appropriate inflection.

French: First Versus Second Language Acquisition

The LAS program had been tested out on a subset of French and a similar subset of English. To establish that ALAS was at least as good a learning system as LAS we wanted to show it capable of acquiring the same language subset. To this end we trained it on the French fragment that had the same syntax as that given to LAS. An example of an input to the program is:[2]

LE MOYENNE ROMBE EST APRES D'UN CROIX JAUNE QUI EST AU-DESSUS D'UN GRAND PENTAGONE NOIR
(BEHIND (KNOW (MEDIUM (DIAMOND A))) (NOT KNOW (YELLOW (CROSS D)) (ABOVE A (NOT KNOW (LARGE (BLACK (PENTAGONE E)))))))
Translation: The medium diamond is behind a yellow cross that is above a large black pentagon.

This sytem worked with a somewhat larger vocabulary than the LAS system consisting of six prepositions, eight nouns, six colors, and three sizes. Not wanting to have to sit through many hundreds of training trials I decided to run this simulation with pause information that would enable it to properly associate its function morphemes like *le* with morphemes already assigned meaning like *carre*.

This example brings up a couple of interesting issues of meaning representation. The first has to do with the semantic correlates of the choice between definite and indefinite articles. It is assumed that definite objects are flagged as known and indefinite objects are flagged as not known. While this is certainly part of what controls the choice it is clear that other things are involved. Thus, this learning simulation solves but a fraction of the issue of article selection. The learning principles may be capable of dealing with the full complexity of article use, but we did not present to the program rich enough input to permit the induction.

The second representational issue concerns the semantic structure of noun phrases. It may be a linguistic universal (Clark & Clark, 1977) that adjective modifiers in noun phrases organize around (before or after) the noun with the more noun-like adjective closer. In our semantic representation we have the adjective predicates so organized around the noun (i.e., color closer than size). This amounts to the claim that the universal tendency in adjective ordering reflects a universal of cognition. It would be possible for ALAS to learn any specific sequence of adjectives, but reasonably enough, ALAS could not learn the "nouniness" principle for adjective ordering unless the nouniness property of adjectives were represented. We could make nouniness a property of the adjectives and leave the learning to discrimination, but there is evidence (McWhinney, personal communication) that children's initial multiple-adjective sequences obey the nouniness principle. Therefore, it seemed better to have the nouniness ordering directly reflected in the structure of the semantic referent.

---

[2]Certain morphemes like *au* are hyphenated to the content words. This is a feature not critical for the success of ALAS but was critical for LAS.

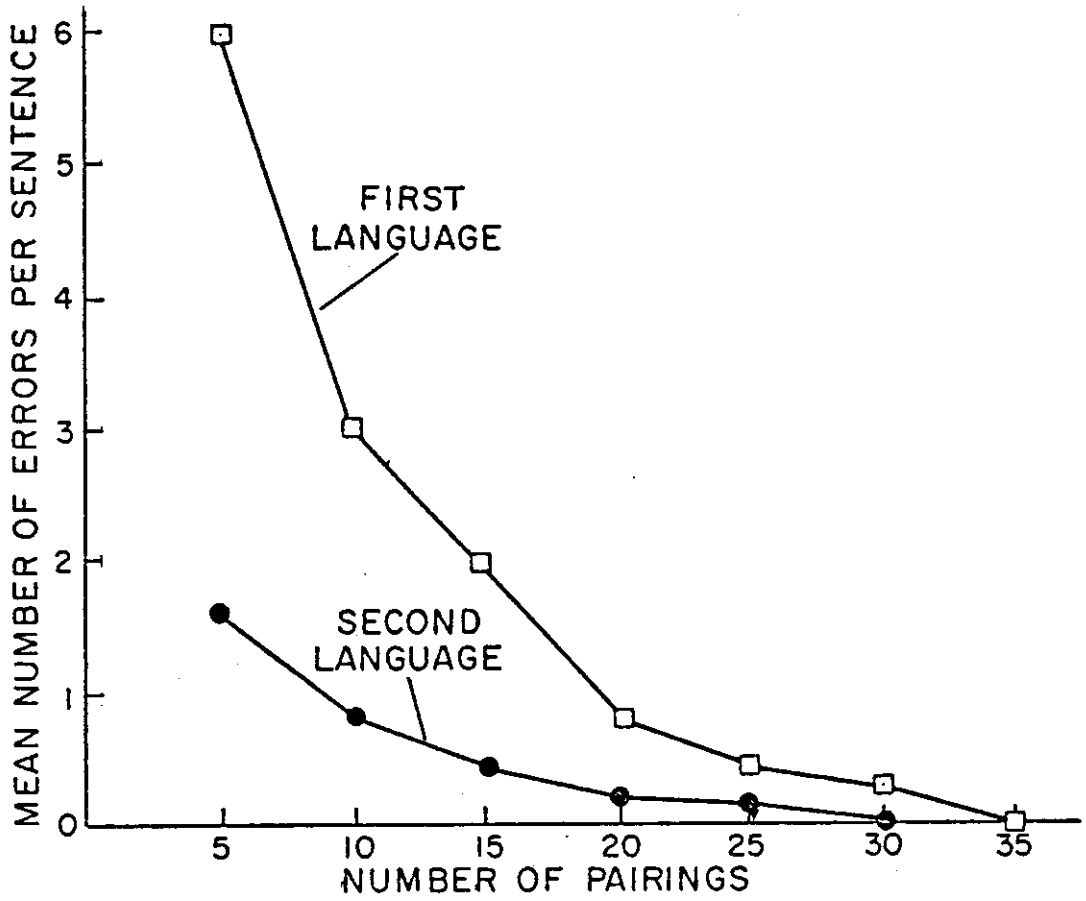

Figure 5

Figure 5 illustrates the rate of learning in this (the first language) condition and another (the second language) condition to be explained. Mean number of errors per sentence are plotted, averaged for blocks of five sentences. After 35 pairings, ALAS had converged on a grammar adequate to deal with this language subset. The rate of learning is considerably more rapid for this language subset than the LATIN subset despite the fact that the grammar we used for French generates an infinite number of constructions (because of relative clause recursion) whereas the Latin grammar we used only generates a finite (albeit > 7 million) sentences. The learning is more rapid in this example because the context-free rules formulated for the French sample do not need as much discrimination as those for the LATIN sample. One of the interesting discriminations that ALAS had to make to learn this subset involved adjectives. Sizes precede the noun while colors followed. ALAS learned to make this discrimination on the basis of the class properties, *color* and *size*. Below are given some examples of ALAS generations and target sentences at various moments in the learning history.

Sentence 6: A-GAUCHE UN CERCLE DU PENTAGONE DEVANT DU TRIANGLE VERT
vs. UN CERCLE EST A-GAUCHE DU PENTAGONE QUI EST DEVANT DU TRIANGLE QUI EST VERT

Sentence 12: UN PETIT ETOILE ROUGE
vs. UN PETIT ETOILE EST ROUGE

Sentence 18: UN PETIT PENTAGONE ROUGE EST DEVANT DU CARRE
vs. UN PETIT PENTAGONE QUI EST ROUGE EST DEVANT DU CARRE

Sentence 27: LE OVALE QUI EST DEVANT D'UN PETIT PENTAGONE ROUGE EST VERT
vs. LE OVALE QUI EST DEVANT D'UN PETIT PENTAGONE ROUGE EST VERT

I was interested in what would happen if instead of using the standard semantic structure as input into the language generation I used strings from a second language bracketed so as to indicate their surface structure. So in another simulation I provided the program with pairings such as:

UN GRAND CARRE QUI EST VERT EST PETIT

((A (LARGE (SQUARE)) (THAT IS GREEN)) IS SMALL)

I view this as an instance of second language acquisition where the learner is mapping from strings of his first language. As can be seen from Figure 5 the learning proceeds even more rapidly. The reason for this is that the word order of French is much more similar to the word order of English than it is to the order of elements in the semantic referent. Therefore, many times the default rule in ALAS worked out, simply mapping order in the English referent into order in the utterance. Therefore, in many cases there was nothing to learn. Presumably, if we took as the first language a language with a very different syntactic structure than French, then it would have been harder to learn French than if we started with the semantic referent. Thus, ALAS reproduces another well-worn observation about language acquisition: Children learning their first language find all languages about equally difficult (assuming they are all of approximately equal similarity to the semantic structure). Adults, learning a new language, experience a great range of difficulty depending on the target language.

Table 1

SAS Schema

Background

S1 is a side of ΔXYZ
S2 is a side of ΔXYZ
A1 is an angle of ΔXYZ
A1 is included by S1 and S2
S3 is a side of ΔUVW
S4 is a side of ΔUVW
A2 is an angle of ΔUVW
A2 is included by S3 and S4

Hypothesis

S1 is congruent to S3
S2 is congruent to S4
A1 is congruent to A2

Conclusion

ΔXYZ is congruent to ΔUVW

Comment

This is the side-angle-side postulate

Verb Auxiliaries

The third simulation was an attempt to have ALAS learn the verb auxiliary system of English. This is one of the standard language fragments used to introduce and motivate transformational grammar (e.g., Culicover, 1976). This is interesting because the verb auxiliary system does not involve any violations of the graph deformation condition and should be learnable by ALAS without resorting to transformations. Table 1 characterizes the set of sentences that we sampled from and presented to ALAS. Although not indicated there, the sentences did, of course, have subject-verb number agreement. The modals we used were *can, could, should, would, will,* and *may* with corresponding meaning components of present-able, past-able, obligation, intention, future, and possibility. These meaning components were not assigned to the terms but rather had to be learned from context. We used sets of four adjectives, eight nouns, six transitive verbs, and four intransitive verbs. Among the verbs were *hit, shoot,* and *run* which all have irregular inflections. Therefore, another problem for the simulation will be to learn the special inflections associated with these terms. As in the French example we provided these strings with the pause structures to permit segmentation.

As can be seen by inspecting Table 1, the meaning structure for the verbs and their auxiliaries is represented as a series of embeddings with modals (and past and present) most external, perfect next, progressive and stative next, and verb most internal. This is analogous to the embedding structure that we set up for nouns. Of these the modal and the verb are obligatory and the remainder optional. While I know of no hard evidence about universality, it does seem that many languages respect this ordering on verb auxiliaries (McWhinney, personal communication).

Figure 6 plots the performance of the system in the first 700 pairings. At the time of this writing we have not yet trained ALAS to perfect performance on this language subset. After 500 trials, it makes a mistake on the auxiliary structure of about one out of four sentences that it generates. I think it is just a matter of time until these errors are corrected. Part of the problem is that there are numerous contingencies to be learned and opportunities to learn each come up rarely. Examples of sentences it generated are:

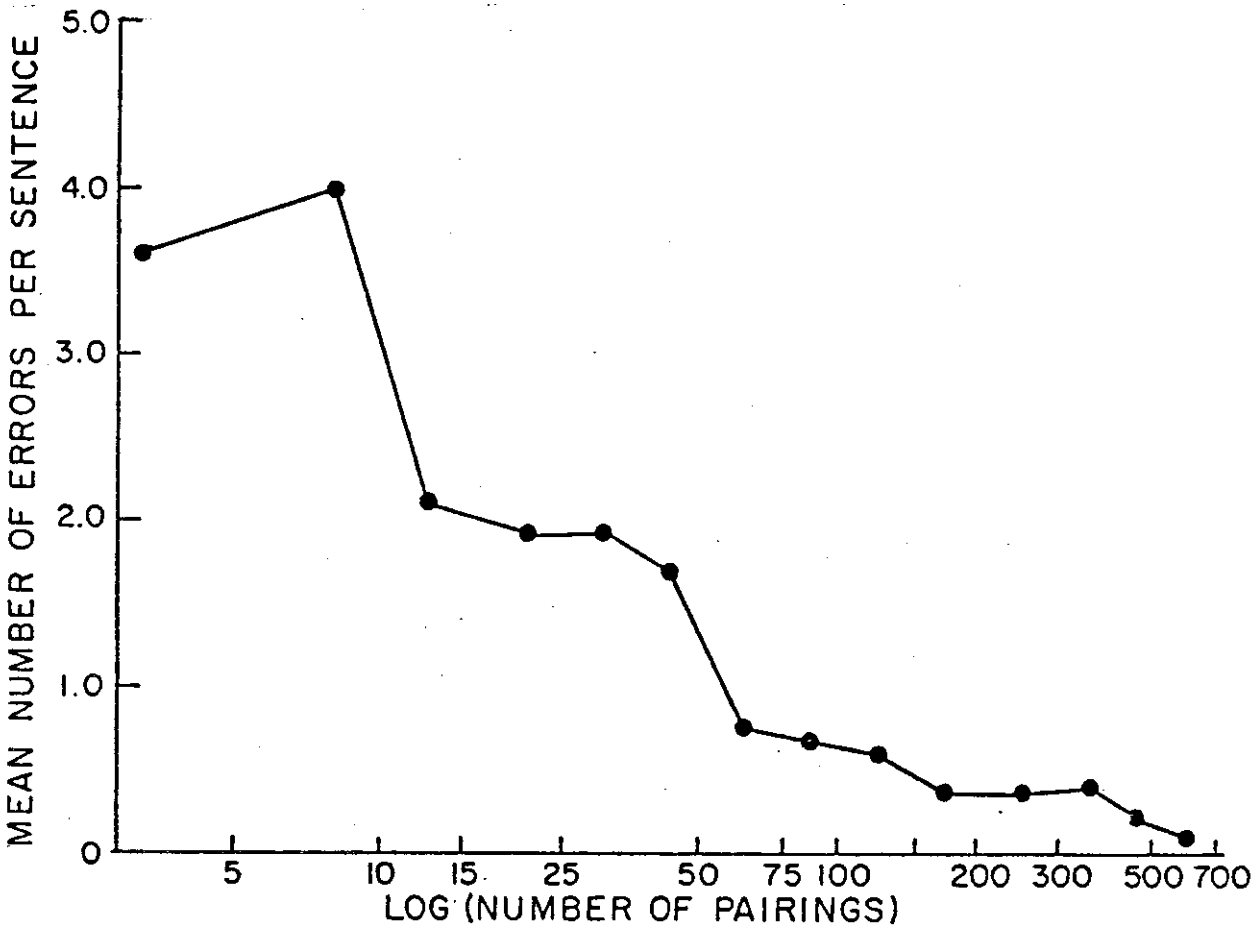

Figure 6

Sentence 1: Jump angry debutante
Sentence 6: Be tickle some actress the sad debutante s
Sentence 10: A tall lawyer s could jump ed
Sentence 16: Some smart actress have tickle ed the sailor s
Sentence 30: Being smart a angry lawyer
Sentence 51: The sailor s were dance ing
Sentence 75: A smart sailor tickle ing a bad lawyer
Sentence 85: The doctor s is been kiss ed by the good hippie
Sentence 110: The bad lawyer should be tickle ing the doctor
Sentence 131: A sailor were was kiss ed by some hippie s
Sentence 148: The farmer may have shoot ed some Arab s
Sentence 174: The actress stab the tall farmer
Sentence 195: The fat doctor s should dance ed
Sentence 213: A fat lawyer can be tall ed
Sentence 228: Some smart lawyer s should be tickle ing the angry actress s
Sentence 253: A sailor are tickle ed by some good lawyer s
Sentence 298: The hippie s would dance ed
Sentence 319: Some hippie s should have been kiss ed by the Arab s
Sentence 354: Some sad ed lawyer s have run
Sentence 370: The sad doctor s are kick ed by the angry farmer s
Sentence 426: Some lawyer s were being hit ed.

These sentences illustrate one of the unexpected developments in the simulation. ALAS collapsed adjectives, transitive verbs, and untransitive verbs into a single word class over time because all these are involved in numerous similar auxiliary structures. This accounts for the appearance of constructions like "sad ed lawyers" and "can be tall ed" where the "ed" inflection has generalized from the verbs to adjectives. Then ALAS had to go through a number of discriminations in which it used the action-quality property distinction between verbs and adjectives to properly restrict the rules.

An important feature of the verb auxiliary system is that, if we consider the verb matrix sequenced tense-modality-perfect-progressive-verb, tense conditions an inflection in the term that immediately follows it, perfect an inflection in the term that follows it, and similarly progressive. This is interesting because the modality, perfect, and progressive terms are all optional. This means that the term inflected for tense or perfect will vary. So, for instance, depending on the verb matrix we inflect perfect (has/had), progressive (is/was), or verb (kicks/kicked) for tense. This is handled in standard transformational analysis by a transformation called affix hopping. This is handled in our simulation by making the prior term part of the rule. So, for instance, ALAS learned the rule:

```
1 + 2 --> 1 + s + 2      if 1 is progressive
                            and the context is present and the syntactic
                            subject is singular
```

It is not a simple matter to judge whether the affix hopping transformation (together with its many support rules) provides a more parsimonious characterization of verb auxiliary structure or whether our context-sensitive rules do. However, the ALAS rules seem much easier to learn. This is one illustration of many where learning considerations can be used to guide linguistic description.

There is one aspect of the slow rate of learning in this simulation that could have been avoided with an extension of the ALAS learning mechanisms. A somewhat interesting example involves the inflection for subject number that controls the distinction between the is-are (and was-were) auxiliary for the stative and progressive markers (which incidentally were collapsed into a single class). ALAS initially found two examples that differed in number of logical subject and formulated the following rule:

```
(1  2) --> 1 + s + 2     if 1 in stative-progressive class
                            and in present context and logical
                            subject is singular
```

However, as illustrated by passive constructions, it is the grammatical subject, not the logical subject that controls verb inflection. Therefore, LAS created a new discrimination to correct the above rule.

```
(1  2) --> 1 + s + 2     if 1 is in stative-progressive class
                            and in present context and logical subject
                            is singular and grammatical subject
                            is singular
```

The problem with this rule is that it is restricted to cases where logical subject is singular and a separate rule must be formed when logical object is plural. This could be avoided if the ALAS

generalization mechanism were called to bear on the output of discrimination and dropped out the "logical subject is singular" feature.

Another problem with the current ALAS simulation is that it is forming separate rules for each different context. Thus, it has a rule for inflection of the verb preceded by *would* and a separate rule for the verb preceded by *should*. Just as it can collapse *would* and *should* into a single class for purposes of rules that generate these terms, so ALAS should treat these terms as classes when they serve as conditions on another generation. This is another example of where the generalization mechanisms should be called on the output of the discrimination process.

The Future

It is clear that ALAS is a considerable improvement over its LAS predecessor and at the time of this report the program is in a state of rapid improvement. The last verb auxiliary example illustrated the need for a more general conception of the generalization process. I would like to try the ALAS system out on other language subsets. The current examples, for instance, did not tap ALAS's facility for learning transformations. I would also like to look at the issue of concept development. It is clear in language acquisition that much of what is controlling syntactic development is development of the appropriate concepts. It also seems likely in cases such as the verb auxiliary system or the definite-indefinite article contrast that efforts to acquire control of the syntax may be part of what is driving the conceptual acquisition. Also, in line with increasing the psychological accuracy of the program to still greater detail, I would like to start to introduce working memory limitations.

References

Anderson, J.R. Language acquisition by computer and child. Technical Report No. 55, Human Performance Center, 1974.

Anderson, J.R. *Language, Memory, and Thought*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

Anderson, J.R. Induction of augmented transition networks. *Cognitive Science*, 1977, *1*, 125-157.

Anderson, J.R. Tuning of search of the problem space for geometry proofs. *IJCAI-81*, submitted.

Anderson, J.R., Greeno, J.G., Kline, P.J., & Neves, D.M. Acquisition of problem-solving skill. In J.R. Anderson (Ed.), *Cognitive Skills and their Acquistion*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

Anderson, J.R. & Kline, P.J. A learning system and its psychological implications. *Proceedings of IJCAI-79*, 1979, 16-21.

Anderson, J.R., Kline, P.J., & Beasley, C.M. Complex learning processes. In R.E. Snow, P.A. Federico, & W.E. Montague (Eds.), *Aptitude, Learning, and Instruction: Cognitive Process Analyses*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Anderson, J.R., Kline, P.J., & Lewis, C. A production system model in language processing. In P. Carpenter & M. Just (Eds.), *Cognitive Processes in Comprehension*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Birnbaum, L. & Selfridge, M. Problems in conceptual analyses of natural language. Research Report # 168, Computer Science Department, Yale University, 1979.

Bresnan, J. A theory of lexical rules and representations. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press, 1981.

Clark, E.V. *First Language Acquisition*. Stanford University, 1975

Clark, H.H. & Clark, E.V. *Psychology and Language: An Introduction to Psycholinguistics*. New York: Harcourt, Brace, Jovanovich, 1977.

Culicover, P.W. *Syntax*. New York: Academic Press, 1976.

de Villiers, J.G. & de Villiers, P.A. *Language Acquisition*. Cambridge, Mass.: Harvard University Press, 1978.

Elio, R. & Anderson, J.R. Effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*. In revision.

Hayes, J.R. & Clark, H.H. Experiments on the segmentation of an artificial speech analogue. In J.R. Hayes (Ed.), *Cognition and the Development of Language*. New York: Wiley, 1970.

Langley, P. Language acquisition through error recovery. Paper presented at the AISB Workshop on Production Systems in Psychology, Sheffield, England, 1981.

Lesser, V.R., Hayes-Roth, F., Birnbaum, M., & Cronk, R. Selection of word islands in the Hearsay II speech understanding system. *Proceedings 1977 IEEE International Conference on ASSP*, Hartford, Ct., 1977.

MacWhinney, B. Basic Syntactic Processes. In S. Kuczaj (Ed.), *Language development: Syntax and Semantics*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Maratsos, M.P. & Chalkley, M.A. The internal languages of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's Language Vol. I*. New York: Gardner Press, 1981.

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

Pinker, S. A theory of the acquisition of lexical-interpretive grammars. In J. Bresnan (Ed.), *The mental representation of grammatical relations*. Cambridge, Mass.: MIT Press, 1981.

Schank, R.C. *Conceptual information processing*. Amsterdam: North Holland, 1975.

Schustack, M.W. Task-dependency in children's use of linguistic rules. Paper presented at the annual meeting of the Psychonomic Society, Phoenix, Arizona, 1979.

Slobin, D.I. Cognitive prerequisites for the development of grammar. In C.A. Ferguson & D.I. Slobin (Eds.), *Studies of child language development*. New York: Holt, Rinehart, & Winston, 1973, 175-208.

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Theory of Language Acquisition Based on General Learning Principles		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) John R. Anderson		8. CONTRACT OR GRANT NUMBER(s) N00014-79-C-0661
9. PERFORMING ORGANIZATION NAME AND ADDRESS Carnegie-Mellon University Schenley Park Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel & Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE June 17, 1981 13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Language Acquisition	Learning	Discrimination
Computer Simulation	Production System	Induction
Skill Acquisition	Problem-Solving	
Syntax	Generalization	
Language Generation		

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A simulation model is described for the acquisition of the control of syntax in language generation. This model makes use of general learning principles and general principles of cognition. Language generation is modelled as a problem solving process involving principally the decomposition of a to-be-communicated semantic structure into a hierarchy of subunits for generation. They syntax of the language controls this decomposition. It is shown how a sentence and semantic structure can be compared to infer

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract (Continued)

the decomposition that led to the sentence. The learning processes involve generalizing rules to classes of words, learning by discrimination the various contextual constraints on a rule application, and a strength process which monitors a rule's history of success and failure. This system is shown to apply to the learning of noun declensions in Latin, relative clause constructions in French, and verb auxiliary structures in English.