In *Categories and Processes in Language Acquisition*, Y. Levy, I. M. Schlesinger and M. D. S. Braine (eds.), Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

7 Learning Syntax and Meanings Through Optimization and Distributional Analysis

J. Gerard Wolff
Praxis Systems plc, Bath, England

INTRODUCTION

It is perhaps misleading to use the word *theory* to describe the view of language acquisition and cognitive development, which is the subject of this chapter. This word is used as a matter of convenience; it applies here to what is best characterized as a partially completed program of research—a jigsaw puzzle in which certain pieces have been positioned with reasonable confidence, while others have been placed tentatively and many have not been placed at all. The most recent exposition of these ideas is developed in two papers: Wolff (1982) and Wolff (1987). Earlier papers in this program of research include Wolff (1975, 1976, 1977, 1980).

Wolff (1982) describes a computer model of linguistic/cognitive development and some associated theory. Wolff (1987) describes extensions to the ideas in the first paper. These papers and previous publications are somewhat narrow in scope, concentrating on detailed discussion of aspects of the theory. The intention here is to provide a broader perspective on the set of ideas.

The chapter begins with a brief summary of the presuppositions of the theory. Then the theory is described in outline: first a brief description of the computer model which is the main subject of Wolff (1982) and then a more abstract account, including the developments described in Wolff (1987). The body of the chapter reviews the empirical support for the theory.

PRESUPPOSITIONS OF THE THEORY

There is space here only for a rather bald statement of theoretical and empirical assumptions on which the theory is based. I will make no attempt to justify these ideas.

1. The theory belongs in the *empiricist* rather than the *nativist* tradition: It seems that language acquisition may very well be a process of abstracting *structure* from linguistic and other sensory inputs where the innate knowledge which the child brings to the task is largely composed of perceptual primitives, structure-abstracting routines, and procedures for analysing and creating language. A *triggering*, nativist view cannot be ruled out *a priori* but the other view is plausible enough to deserve exploration.

2. It seems clear that, while children may be helped by explicit instruction in language forms, by reward for uttering correct forms, by correction of errors, and by other *training* features of their linguistic environment, including the grading of language samples, they probably do not need any of these aids. It seems prudent, as a matter of research strategy, to think in terms of learning processes which can operate without them but which can take advantage of them when they are available.

3. In a similar way it seems prudent to develop a theory in which learning does not depend on prelinguistic communicative interaction between mother and child but which is at the same time compatible with the fact that such interactions clearly do occur.

4. Although semantic knowledge may develop earlier than syntactic knowledge (or make itself apparent to the observer at an earlier age) it seems that the learning of both kinds of knowledge is integrated in a subtle way. One kind of knowledge is not a prerequisite for the learning of the other.

5. Mainly for reasons of parsimony in theorizing, it has been assumed that a uniform set of learning principles may be found to apply across all domains of knowledge—which is not to deny that differences may also exist. The mechanisms proposed in the theory appear to have a wide range of application.

6. It is assumed that there is a core of knowledge which serves both comprehension and production processes. The theory is framed so that the representation of this core knowledge and the posited processes for learning it are broadly compatible with current notions about processes of comprehension and production.

OUTLINE OF THE THEORY

As already indicated, the theory is based on the kinds of empiricist ideas of associationism and distributional analysis which were so heavily criticized by Chomsky (1965). Those earlier ideas have been extended and refined in two main ways:

- A series of computer models have been built and tested to provide detailed insights into the nature of the proposed mechanisms and their adequacy or otherwise to explain observed phenomena.

- The early ideas are now embedded within a broader theoretical perspective: learning may be seen as a process of optimization of cognitive structures for the several functions they must serve.

This section of the chapter will describe the theory in two stages:

1. a relatively concrete description in terms of the most recent of the computer models in which the theory is embodied: program SNPR.
2. a more abstract or “conceptual” view which includes ideas not yet incorporated in any computer model.

Program SNPR

Table 7.1 summarizes the processing performed by the SNPR model. The *sample of language* is a stream of letter symbols or phoneme symbols without any kind of segmentation markers (spaces between words, etc.). The main reason for leaving out all segmentation markers is to explore what can be achieved without them, given that they are not reliably present in natural language.

The letter or phoneme symbols represent *perceptual primitives* and should not be construed as letters or phonemes *per se.* If the model is seen as a model of

TABLE 7.1
Outline of Processing in the SNPR Model

```
1.  Read in a sample of language.

2.  Set up a data structure of elements (grammatical rules) containing, at this stage, only the primitive
    elements of the system.

3.  WHILE there are not enough elements formed, do the following sequence of operations repeatedly:

    BEGIN

        3.1   Using the current structure of elements, parse the language sample, recording the
              frequencies of all pairs of contiguous elements and the frequencies of individual elements.

              During the parsing, monitor the use of PAR elements to gather data for later us in
              rebuilding of elements.

        3.2   When the sample has been parsed, rebuild any elements that require it.

        3.3   Search amongst the current set of elements for shared contexts and fold the data structures
              in the way explained in the text.

        3.4   Generalize the grammatical rules.

        3.5   The most frequent pair of contiguous elements recorded under 3.1 is formed into a single
              new SYN element and added to the data structure. All frequency information is then
              discarded.

    END
```

syntax learning then the symbols may be seen as perceptual primitives like formant ratios and transitions. If the model is seen as a model of the learning of nonsyntactic cognitive structures (discussed later) then the symbols may be seen as standing for analyzers for colors, lines, luminance levels, and the like.

Elements in the data structure are of three main types:

- Minimal (M) elements. These are primitives (ie letter or phoneme symbols).
- Syntagmatic (SYN) elements. These are sequences of elements (SYN, PAR, or M).
- Paradigmatic (PAR) elements. These represent a choice of one and only one amongst a set of two or more elements (SYN, PAR, or M).

The whole data structure has the form of a phrase-structure grammar; each element is a *rule* in the grammar. Although it starts as a set of simple rules corresponding to the set of primitives, it may grow to be an arbitrarily complex combination of primitives, sequencing rules (SYN elements), and selection rules (PAR elements). This grammar controls the parsing process.

The general effect of the repeated application of operations 3.1 (parsing and recording the frequencies of pairs) and 3.5 (concatenation of the most frequent pair of contiguous elements) is to build SYN elements of progressively increasing size. Early structures are typically fragments of words; word fragments are built into words, words into phrases and phrases into sentences.

The effect of operation 3.3 (sometimes called *folding)* is to create *complex* SYN elements, meaning SYN elements which contain PAR elements as constituents. For example, if the current set of elements contains 1 → ABC[1] and 2 → ADC, then a new PAR element is formed: 3 → B | D[2] and the two original SYN elements are replaced by a new SYN element: 4 → A(3)C. Notice that A, B, C, and D may be arbitrarily complex structures. Notice also how the context(s) of any element is defined by the SYN element(s) in which it appears as a constituent.

Operation 3.4 creates *generalizations* by using the newly formed PAR elements. For example, element 3, just described, would replace B or D in other contexts: 5 → EB would become 6 → E(3), and so on. Generalizations may also be produced by operation 3.5 as explained in Wolff (1982).

Operations 3.3 (folding) and 3.4 (generalization) do not come into play until

[1] The notation "1 → ABC" means "the symbol '1' may be rewritten as ABC" or "the symbol '1' is a label for the structure ABC." To aid understanding in this and later examples, integer numbers have been used for references (labels) to structures ("nonterminal symbols" in grammatical jargon), while capital letters are used to represent the material described in the grammar ("terminal symbols").

[2] Read this as "the symbol '2' may be rewritten as B or D."

enough SYN elements have been built up for shared contexts to appear. Likewise, operation 3.2 (rebuilding) will not play a part in the learning process until some (over)generalizations have been formed.

Correction of Overgeneralizations

The *monitoring* and *rebuilding* processes shown in Table 7.1 are designed to solve the problem of *overgeneratizations:* If it is true that children can learn a first language *without explicit error correction* (and there is significant evidence that this is so), how can a child learn to distinguish erroneous overgeneralizations from the many *correct generalizations* that must be retained in his or her cognitive system?

Figure 7. 1 illustrates the problem. The smallest envelope represents the finite, albeit large, sample of language on which a child's learning is based. The middle sized envelope represents the (infinite) set of utterances in the language being learned. The largest envelope represents the even larger infinite set of all possible utterances. The difference between the middle sized envelope and the largest one is the set of all utterances which are not in the language being learned.

To learn the language, the child must generalize from the sample to the language without overgeneralizing into the area of utterances which are not in the language. *What makes the problem tricky is that both kinds of generalization, by definition, have zero frequency in the child's experience.*

Notice in Fig. 7. 1 that the sample from which the child learns actually overlaps the area of utterances not in the language. This area of overlap, marked

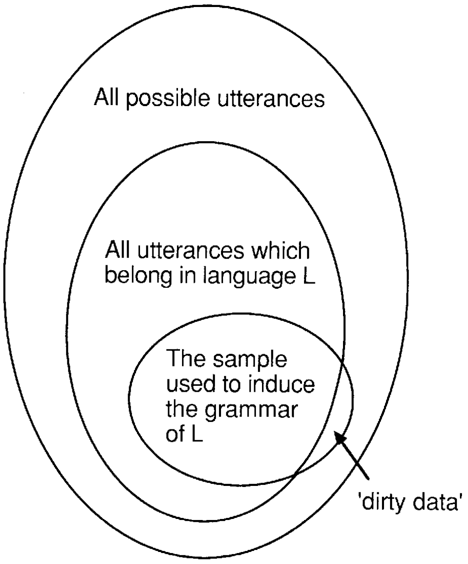

All possible utterances
All utterances which belong in language L
The sample used to induce the grammar of L
'dirty data'

FIG. 7.1. Kinds of utterance in language learning.

‘dirty data’, and the associated problem for the learning system, is discussed later in the chapter.

To correct overgeneralizations, the monitoring process in SNPR keeps track of the usage of all constituents of all PAR elements in all the contexts in which they occur (remember that contexts are defined in terms of the elements built by SNPR). If any PAR element fails to use all its constituents in any context then it is *rebuilt* for that context (and only that context) so that the unused constituent(s) is removed. As a hypothetical example, a PAR element 1 → P | Q | R may fail to use R in the context 2 → A(1)B. In such a case it becomes 3 → P | Q and 2 is rebuilt as 2 → A(3)B. The structure 1 → P | Q | R may still be used in other contexts.

This mechanism, in which structures are eliminated if they fail to occur in a given context within a finite sample of language, is an approximation to what one imagines is a more realistic mechanism which would allow the *strength* of structures to vary with their contextual probability.

This kind of mechanism will allow a child to observe that “mouses,” for example, is vanishingly rare in adult speech and will cause the speech pattern for “mous” to be removed (or *weakened)* in the structure which generates “mouses,” “houses,” “roses,” etc. The correct form (“mice”) will be learned independently.

Preserving Correct Generalizations. What is special about the mechanism in SNPR for correcting overgeneralizations is that certain kinds of generalization cannot be removed by it. The mechanism thus offers an explanation of how children can differentiate *correct* and *incorrect* generalizations without explicit error correction.

To see how it is that the rebuilding mechanism cannot touch some generalizations, consider the following example. From a text containing these three sentences:

John sings
Mary sings
John dances

it is possible to induce a fragment of grammar like this:

1 → (2)(3)
2 → John | Mary
3 → sings | dances

Notice that there is a generalization: the grammar generates “Mary dances” even though this was not in the original sample.

Notice, in particular, that the monitoring and rebuilding mechanism cannot remove this generalization. The reason is that, in the sample from which the grammar was induced, “sings,” “dances,” “John,” and “Mary” are *all* used in the context of the structure “1.”

In running SNPR, many examples have been observed like this where generalizations are preserved and distinguished from other generalizations which are eliminated.

Other Mechanisms. There is no space here for a full discussion of the problem of correcting overgeneralizations without external error correction. The mechanisms in SNPR are one of only a few proposals that have been put forward. Braine (1971) has proposed a mechanism but I have not been able to understand from the description how it can remove overgeneralizations without at the same time eliminating correct generalizations. The proposal by Coulon & Kayser (1978) apparently fails because, judging by the sample results they give, wrong generalizations are allowed through the net. The “discrimination” mechanism in Anderson (1981) seems to depend on the provision of explicit *negative* information to the model.

Other mechanisms that have been proposed (e.g., Langley, 1982) use covert error correction; to do this they need to make what I believe are unwarranted assumptions:

- that a child’s knowledge of meanings may be used to correct overgeneralizations. If the learning process has to bootstrap semantic structures as well as syntactic structures (as children apparently do), then some other mechanism is needed for the correction of overgeneralizations.
- that there is a one-to-one relation between syntax and meanings. This is quite clearly false for natural language.

The process in SNPR depends on *relative contextual probabilities* of structures and does not employ any notion of falsification of hypotheses or the like.

The notion that a child may learn by creating hypotheses and observing whether they are confirmed or falsified is unsound for much the same reasons that scientific hypotheses cannot be either confirmed or falsified (Lakatos, 1978). A full discussion of this interesting issue is not possible here.

Summary

To summarise this outline of SNPR’s functioning, the overall behaviour of the program is to build up cognitive structures by concatenation, using frequency as a heuristic to select appropriate structures. Interwoven with the building process are processes to form disjunctive groups, to form generalizations and to correct overgeneralizations. The structures built by the program have the form of unaugmented phrase structure grammars.

The Abstract View of the Theory

Program SNPR embodies most but not all of the ideas in the theory. This section describes the model in more abstract terms than in the previous section and incorporates ideas from the most recent phase of research (described in Wolff, 1987).

Taking the abstract view, the central idea in the theory is that language acquisition and other areas of cognitive development are, in large part, processes of building cognitive structures which are in some sense *optimal* for the several functions they have to perform. This view is a development of notions of "cognitive economy" which were in vogue in the 1950s and which have attracted intermittent attention subsequently.

This abstract view fits well with and in a sense grows from a recognition that human cognitive systems are products of natural selection and are therefore likely to be conditioned by principles of efficiency.

Compressing Cognitive Structures

One of the functions of a cognitive system is to be a store of knowledge. It seems clear that, *other things being equal,* storage demands of cognitive structures should be minimized. The brain's storage capacity is large, no doubt, but it is not infinite and it seems reasonable to suppose that natural selection would have favored compact storage.

There are at least six ways of reducing the storage demands of a body of data:

1. A pattern (a sequence of elements) which is repeated in a variety of contexts may be stored just once and then accessed via pointers from several contexts. A sequence like

ABCDPQRABCDABCDPQRABCDPQRPQR

may be reduced to 12112122, where 1 → ABCD and 2 → PQR.

This is *chunking.* It is also like the use of subroutines in computer programs.

2. Two or more patterns sharing the same context may be placed in a disjunctive group which is accessed via a reference or pointer from the given context. This saves repeated storage of the context pattern. For example, ABCPQRDEF and ABCXYZDEF may be reduced to ABC(1)DEF where 1 → PQR | XYZ. This is *folding.*

3. *Frequent, large* patterns in 1 will clearly produce a bigger saving than rare small patterns. For reasons spelled out in Wolff (1982) it is best to concentrate on frequency in searching for repeating patterns. There is here a clear theoretical justification for regarding frequency as an important variable in learning. The importance of frequency was recognized in associationist psychology (e.g., Carr, 1931), mainly for intuitive reasons, but it fell out of favor when associationism went out of fashion.

4. Repeating contiguous instances of a pattern may be recorded just once and marked for repetition. For example, AAAAAAAAAAA may be reduced to A* or A[11].[3]

This is *iteration.* A device with similar effect is *recursion.*

5. Storage space may be saved by simple *deletion* of information or *not recording it.* As discussed in Wolff (1982), there is a close connection between this mode of economy and the phenomenon of *generalization.* This point is amplified below in discussing the tradeoff between the size of a knowledge structure and its usefulness.

6. The last technique in this list is the principle of *schema-plus-correction;* this is described and discussed in Wolff (1987). The idea here, of course, is that a pattern may be recorded by reference to a class or schema of which it is an example, together with the details (corrections) that are specific to the given item. “Tibs” may be described as “cat[tabby, 5-years-old, one-leg-missing].”

Five of these techniques for reducing storage (or transmission) costs of data (items 1, 2, 3, 4, and 6) may be described as techniques for *data compression:* They exploit any *redundancy* that may exist in a body of data; in general, they do not result in a loss of information. The fifth principle is different because information is lost or never recorded.

Using Cognitive Structures as Codes

A second major function of a cognitive system is to provide a set of codes for patterns of information: afferent patterns coming from the senses, efferent patterns transmitted to the organism’s motor system, information patterns transmitted in the course of the brain’s internal data manipulations (i.e., thinking), and also the patterns of information to be stored in the cognitive system itself.

An obvious and relatively simple example is the use of words as codes for perceptual / conceptual complexes; words certainly are not the only codes employed in the nervous system and need not, of course, be employed at all. As before, we are assuming that, *other things being equal,* codes that minimize the amount of information to be used for a given purpose will be preferred over other less efficient codes. Precisely the same compression principles may be applied to minimizing required storage space and maximizing the efficiency of codes.

Tradeoff

There is a tradeoff between the *size* of a knowledge structure and its *power* for encoding knowledge. At one extreme there is a very compact grammar like this:

1 → 2*
2 → A | B | C | … | Z.

[3] “*” means “repeat as many times as desired.”

This small grammar generates any (alphabetic) text of any size; but it achieves no compression because the text is encoded in the conventional way as a stream of characters.

At the other extreme is a "grammar" with one rule like this:

1 → "the complete sample of language observed to date"

This grammar is not at all compact but it provides a very efficient code: Given the existence of the grammar, one small reference ("1") may be used to represent the whole sample.

Between these two extremes lies a spectrum of grammars.

It is perhaps useful to remark in passing that there is a close connection between this spectrum and the phenomenon of *generalization.* The first grammar, above, is extremely general; the second is extremely specific. In between are grammars which, in varying degrees, are more general than any specific language sample.

The connection between generality in grammars and the size / power tradeoff is simple and direct when grammars are unambiguous (when any given pattern can be generated in only one way). The connection is less direct when, as is usually the case with natural languages, grammars are ambiguous.

Learning

In this theory of learning, it is assumed that a child starts with a small very general *grammar* like the first one above and gradually extends it. As the grammar is extended, it will become progressively more *powerful* as a means of encoding information. The term *grammar* in this context is shorthand for "syntactic / semantic structure."

Whether or not it becomes less general at the same time depends on whether the original general rules are retained in the grammar or discarded. They are almost certainly retained *in reserve,* so to speak, for occasions when generality is needed.

Additions to the set of rules should not be made indiscriminately. At every stage, it is likely that the child will choose the more *useful* rules in preference to less useful or powerful rules. According to the theory, the child should, at all times, try to maximize the effectiveness of each new rule as a means of encoding data economically: He or she should try to maximize the *compression capacity* (CC) or descriptive *power* of the grammar. At the same time, the child should try to minimize the *size* of the grammar which is being built; the size of the grammar is termed S_g.

The ratio between the descriptive power of the grammar and its size (i.e., CC/S_g—which is termed the *efficiency* of the grammar) should at all stages be maximized.

It is in the nature of the search process that the most powerful elements (those giving a large increase in CC for a relatively small increase in S_g) will be found early, and progressively less powerful elements will be found as learning proceeds. There is no reason to suppose that learning will cease when CC/S_g reaches a maximum. Rather, we may suppose that learning will cease when candidate elements that the child discovers or constructs do not add enough to the grammar's CC to justify the attendant increase in S_g. This point will depend on the relative value to a given child of CC and the information storage space corresponding to S_g. These values may depend on motivational factors and on the total available storage space among other things; they are likely to vary from one child to another.

The foregoing ideas are illustrated diagrammatically in Fig. 7.2. The graphs show the trade-off between the coding capacity or power of a grammar and its size. In general, big grammars are more powerful than little ones. Independent of this trade-off is the efficiency of a grammar, meaning the ratio of power to size. The most efficient grammars lie along the line marked "1"; the least efficient grammars lie along the line marked "4", and intermediate grammars lie in between these lines.

The learning process starts near the bottom left of the diagram in the region of small grammars which are not powerful. The learning process gradually builds the grammatical system keeping as close as possible to the highest of the lines ("1 ") in the diagram, thus maintaining as much efficiency in the grammar as possible at all times.

Other Factors. Only the two functions mentioned have so far been considered in any detail but it is clear that at some stage others will need consideration. An obvious candidate is the facility with which information may be retrieved from a knowledge structure. As with information transmission, it seems that economy in storage may sometimes be bought at the cost of cumbersome retrieval. Likewise, reliability of cognitive operations may demand the preservation of some redundancy in knowledge structures. A point worth stressing here is

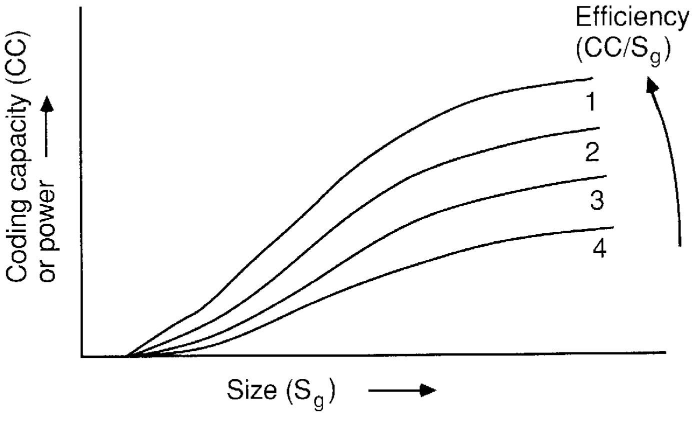


FIG. 7.2. A search space of grammatical systems.

that concepts like *efficiency* and *optimization* are *functional* notions: It is meaningless to say that something is efficient unless one can say what ends are efficiently served. The theory thus offers a bridge between the cognitive and motivational aspects of mental life.

Realization of the Abstract Principles in SNPR

Program SNPR appears to be a realization of the first five of the optimization principles which have been described. No doubt, other realizations are possible. A model incorporating the sixth principle has not yet been attempted.

Constructs like CC and S_g are not employed explicitly by the procedures in SNPR. The effects to be described appear to be *emergent* properties of the SNPR algorithm.

The SNPR model builds its knowledge structures from an initial small base. The effect of the building operation is apparently to increase progressively the CC of the structures while maintaining a high efficiency. The key to this building process is the use of frequency as a heuristic to discriminate *good* structures from *bad* ones. The general tendency of the building mechanisms is to maximize the sum of the products of frequency and size of the structures being built; this promotes a high efficiency.

The generalization mechanisms usually have the effect of reducing S_g without a corresponding reduction in CC; the overall effect is thus usually an increase in efficiency. The rebuilding mechanism apparently has the effect of increasing CC for a given S_g and thus promotes a high efficiency. In general, the mechanisms which have been shown to succeed in discovering a grammar from a sample of its language appear also to be mechanisms that promote high descriptive efficiency in the knowledge structures.

Other Concepts

Before leaving this outline description of the theory and the computer model in which some aspects of it are embodied, we may note some general points about the character of the theory. It gives expression to a number of rather potent ideas, most of which have a fairly long history in psychology and cognitive science.

One of these ideas is the notion that the acquisition of one skill may be a *stepping stone* or foundation for the acquisition of another. This principle is exemplified in SNPR in the way that the program builds a heterarchy of elements, corresponding to a heterarchy of language skills, with complex elements constructed from simpler constituents.

Another useful principle, more recent in origin but now widely recognized, is the idea that knowledge structures may with advantage be constructed from *discrete modules* together with a *restriction on the range of module types* allowed. This feature, realized in the theory by the three types of element, has the advantage that it facilitates the processes of building or modifying a knowledge structure much in the same way that modularity facilitates construction and repair of buildings or electronic systems (or Lego models).

An idea in the theory which seems not to be widely recognized is the *separation of conjunctive groupings and disjunctive groupings* into distinct modules. The significance of this idea depends on a recognition of the significance of conjunction and disjunction in data compression and optimization. The conjoining of elements represents a reduction of *choice* in the knowledge system and thus an increase in its information content *(information* being used here in the Shannon-Weaver sense). Conversely, the disjunctive grouping of elements represents a preservation of choice and a corresponding reduction of information content. The quest for an optimum balance between S_g and CC is facilitated if the groupings which, so-to-speak, pull in one direction are kept separate from the groupings which have the opposite tendency. This design feature of a knowledge system facilitates the process of molding the structure to fit accurately the patterns of redundancy in the data base.

Mention may be made, finally, of a fourth idea, not new, which appears to have a broad significance in psychology and elsewhere. A modular knowledge structure can be optimized by processes akin to the processes of *natural selection* operating in the evolution of animals and plants. Those modules that are *useful* can be allowed to survive while the many rival modules that are less useful or, in some sense, less efficient may be progressively eliminated. This kind of *evolutionary principle* can be seen to operate in SNPR in the way that absolute and contextual frequency governs the retention or elimination of elements.

EMPIRICAL EVIDENCE

Although the theory is by no means fully developed, it is substantial enough for one to ask how well it fits with available data on people's mature knowledge of language and on the developmental processes leading to the mature system. Most of the evidence I review has been presented piecemeal in previous publications; the intention here is to summarize relevant evidence and to expand the discussion of certain points not previously considered in any detail.

Part of the empirical support for the theory lies in the presuppositions discussed above. To the extent that these presuppositions derive from observations (and in this they vary a good deal) the theory is likewise supported.

A second kind of empirical support is provided by the observed phenomena which the theory was designed to explain. To the extent that the theory does demonstrably succeed in providing explanations for these phenomena, they constitute validating data.

There is, lastly, a kind of empirical support, not always very distinct from the other two, which is phenomena not directly addressed by the theory which do nonetheless turn out to be explicable in terms of the theory; this kind of explanatory bonus can be quite persuasive. There are quite a few phenomena in this category that are considered after a review of those observations which the theory was originally designed to explain.

This section on empirical evidence ends with a discussion of certain observations that appear to be incompatible with the theory in its present form.

Phenomena Addressed by the Theory

The main phenomenon addressed by the theory is the observation that children can apparently discover a generative grammar from a sample of language, given only that sample as data. Insofar as SNPR does broadly the same thing, albeit with simpler grammars, it may be regarded as empirically valid. As an example, SNPR has successfully retrieved the grammar shown in Table 7.2, given only a sample of the corresponding language as input (Wolff, 1982).

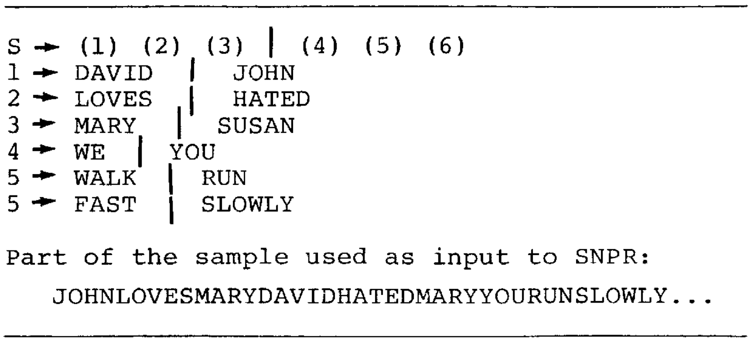
In the following subsections, the components of this grammar-abstraction process are examined individually to see how well the theory fairs in each domain.

1. Segmentation

The first subproblem chosen for this project was to find a sufficient mechanism to explain how children could learn the segmental structure of language given the apparently insufficient and unreliable nature of clues like pause and stress.

The problem was artificially purified by assuming, contrary to probable fact, that such clues made no contribution to the segmentation process. The main alternative is some kind of distributional or cluster analysis designed to reveal statistical discontinuities at the boundaries of words and other segments. Ideas of this kind were, of course, central to distributional linguistics and had been explored by linguists (Gammon, 1969; Harris, 1961, 1970) developing tools for

TABLE 7.2
Artificial Grammar and Fragment of a Corresponding language Sample



```
S → (1) (2) (3) | (4) (5) (6)
1 → DAVID  |  JOHN
2 → LOVES  |  HATED
3 → MARY   |  SUSAN
4 → WE | YOU
5 → WALK  |  RUN
5 → FAST  |  SLOWLY
```

Part of the sample used as input to SNPR:

JOHNLOVESMARYDAVIDHATEDMARYYOURUNSLOWLY...

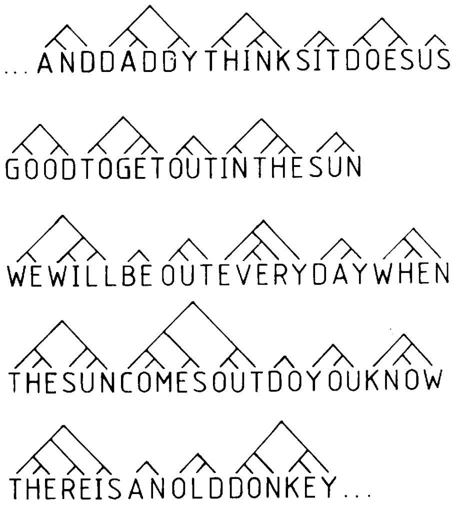


FIG. 7.3. Part of a 10,000 letter sample from book 8A of the Ladybird Reading Series showing a parsing developed by program MK1O at a late stage of processing (Wolff, 1977).

linguistic analysis and also by psychologists (e.g., Olivier, 1968) interested in psychological processes.

Word Structure. After a good deal of experimentation, a program was developed (a precursor of program SNPR called MK10) which produced good results in discovering word structure in artificial and natural language texts (Wolff, 1975, 1977).

A variety of search heuristics were tried, including transition probabilities between elements and measures derived from standard indices of correlation, but the best results by far were obtained with a simple measure of conjoint frequency of elements. In terms of the compression principles (which were recognized after MK10 was developed) this model may be seen as a fairly direct expression of principles 1 and 3.

Figure 7.3 shows part of a sample of an unsegmented text taken from book 8A of the Ladybird reading scheme; the tree markers show the parsing developed by the program at a late stage of processing.

There is an extremely good fit between these markers and the conventionally recognized word structure of the text, showing clearly that the program is sensitive to structures at this level. There is some evidence that the process is also sensitive to structures smaller than words. The performance of the program in identifying structures larger than words is considered in the next section.

Phrase Structure. If program MK10 is run on a text like the one just described, it will, given time, build up structures which are larger than words and which look like phrase-structure trees. The results obtained with the Ladybird text, and others, showed a rather poor correspondence between these trees and the trees which would be assigned to the texts in conventional surface structure analyses.

One possible reason for this poor performance is that the program was not designed to discover disjunctive groupings of elements. Program SNPR does seek disjunctive groupings but it is not yet efficient enough to be run far enough on natural language for its performance with phrase structures to be judged.

A stop-gap solution to the problem of disjunctive relations was to transcribe a text as a sequence of word classes and to use this transcribed text as data for MK10. This is not a wholly satisfactory procedure because it does not provide for disjunctive groupings above and below the level of words. Despite this shortcoming and the other clear shortcomings of MK10 (not taking account of semantics, for example) surprisingly good results were obtained (Wolff, 1980).

Figure 7.4 shows one sentence (and a bit) from a 7600 word sample from Margaret Drabble's novel *Jerusalem the Golden,* which was transcribed as a sequence of word class symbols and processed in that form by MK10. The dendrogram above the sentence shows a supposedly uncontroversial surface structure analysis assigned by a linguist and the author. The dendrograms beneath show the parsing developed by the program at a late stage or processing. There is quite a good correspondence between the two analyses in this and many other cases. Statistical tests have confirmed that the correspondence is very unlikely to be an artifact of chance coincidences. As we have seen, these results on segmentation cannot be construed as proof that children actually do distributional analyses of this kind. They merely demonstrate that such processes are plausible candidates, perhaps sufficient by themselves to explain how children learn to segment language or perhaps working in conjunction with processes which use available prosodic and semantic cues. It is perhaps worth observing that the use of such cues as a guide to structure is, in a deep sense, also distributional. If redundancy and structure are in some sense

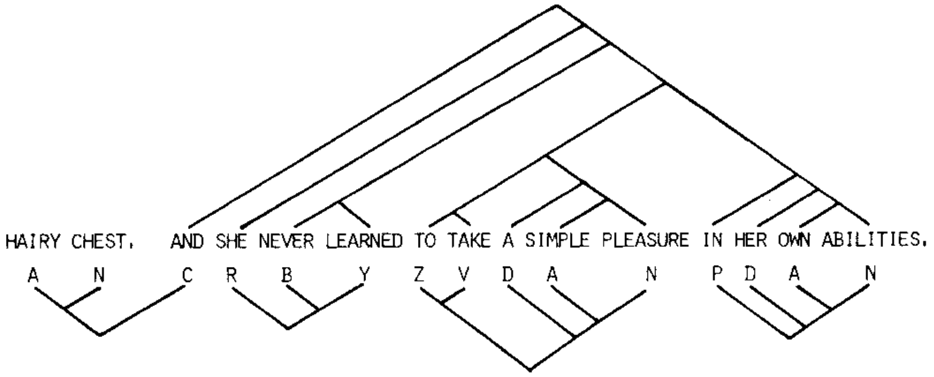

HAIRY CHEST, AND SHE NEVER LEARNED TO TAKE A SIMPLE PLEASURE IN HER OWN ABILITIES.
A N C R B Y Z V D A N P D A N

FIG. 7.4. One sentence from a 7600 word sample from *Jerusalem the Golden* (Margaret Drabble) showing (above the text) a surface-structure analysis assigned intuitively and (below the text) the parsing developed by program MK10 at a late stage of processing (Wolff, 1980). This figure is reproduced by kind permission of Kingston Press Services Ltd.

equivalent (Garner, 1974) then all modes of discovering structure may ultimately be seen in terms of redundancy abstraction.

2. Parts of Speech and Other Disjunctive Categories

In the same way that the theory has developed ideas from taxonomic linguistics about the discovery of segmental structure, it has adopted and extended what is perhaps the most distinctive idea from this tradition: how part-of-speech and other disjunctive categories are established.

The basic idea of course is to look for groups of elements where the members of the group share one or more contexts (where *context* means either or both of syntactic and semantic context). If, for example, the child finds the two patterns AX and BX in the language that he hears then X may be treated as a *context* and a structure (1)X may be created where 1 is a pointer or reference to the disjunctive grouping (A | B).

The principle is quite simple but its proper realization in a fully specified working system has proved quite difficult. What has been achieved in SNPR is the precise specification of a process in which the searches for segmental and disjunctive groupings are *integrated* in such a way that elements of one type may be incorporated in elements of the other kind and this at any level. The discovery procedure produces a generative grammar rather than some less explicit description of the data.

The observations that languages contain disjunctive categories like nouns, verbs, and adjectives is, like any other observation, partly a product of one’s theoretical preconceptions. There are, no doubt, other descriptive frameworks one could employ which do not use them. But categories like these seem to be so strongly determined by the linguistic data that it seems reasonable to characterize them as observed phenomena rather than theoretical constructs. Less well supported but still reasonably clear is the observation that categories like these are a (usually unconscious) part of every adult’s unschooled knowledge of language structure. The main evidence for this derives from word association tests (e.g., Deese, 1965; but see the discussion below on the “S-P shift”) and from speech errors (e.g., Fromkin, 1973).

If it is accepted that there is indeed something here requiring explanation we may ask how well the theory does in this respect. The performance of SNPR with artificial texts gives some indication of its ability to find disjunctive groups but, given that these groups have been artificially created, we cannot tell directly how it would do with natural categories.

Some attempt has been made to run SNPR on natural language but it requires impractically long program runs to get useful results. (This in itself should not be an objection to the model given that children, with much more computational power at their disposal, take several years to develop their linguistic knowledge.) Nevertheless, the program does develop some categories which correspond fairly well with recognized categories in English.

Validation of this general approach, though not the precise details of the current model, is provided in an interesting study by Kiss (1973). (Rosenfeld, Huang, & Schneider, 1969, obtained similar results although their theoretical interests were rather different.)

Using a rather simple definition of the *context* of a word (the word immediately preceding the given word and the word following), Kiss measured the extent to which each of a set of selected words in a sample of natural language shared contexts with other members of the set. He then applied a standard clustering algorithm to these data to determine the *strength* of association between words. The clusters of words identified in this way corresponded quite well with the categories conventionally recognized by linguists (nouns, verbs, etc.).

Mention may be made here of observations which provide some supporting evidence for the idea that children do do a systematic comparison of linguistic structures in an attempt to find elements shared by more than one structure, much as in SNPR. In Ruth Weir's classic study (1962) of her young son's presleep soliloquies one may find sequences of the child's utterances in which a word or a group of words recurs:

> *… which one; two; one; right one; now left one; this one. . .* (p.180)

and later,

> *… I'm taking the yellow blanket; too much; I have the yellow blanket; down; don't stop in the blanket. . .* (p.181)

and many similar examples.

One gets the impression (supported by direct evidence appearing elsewhere in Weir's protocols) that the child is repeating bits and pieces of language heard during the preceding day. The recurrence of words like "one" and "blanket" may simply reflect their recurrence in the original sequence of adult utterances but the overall impression one gets from Weir's records is that utterances are being brought together from disparate sources on the strength of shared constituents. We may here be witnessing part of the process of sorting and sifting required to establish disjunctive groupings of distributionally equivalent elements.

3. Generalization of Grammatical Rules and Correction of Overgeneralizations

The theory (in common with a number of other artificial intelligence theories of language acquisition) provides for the generalization of linguistic rules. Something like this is essential in any (empiricist) theory of language acquisition in order to explain how it is that both children and adults produce novel constructions which they are unlikely ever to have heard.

There is no great difficulty in creating generalizations. Almost any distortion in a grammar, including the deletion of rules, will lead it to generate constructions which it did not generate before. No strong claim is made about the particular generalization mechanisms in the present theory—it is a matter for future investigation to establish what mechanism or mechanisms children and adults actually use.

Correction of Overgeneralizations. The much more difficult problem, which has received relatively little attention from psychologists or other theorists, is to establish what theoretically well-motivated process can, without the aid of a teacher or informant, eliminate the wrong generalizations and retain the good ones as permanent fixtures in the grammar.

The monitoring and rebuilding mechanisms in SNPR offer a possible explanation. Other possibilities were briefly discussed in the section outlining the workings of SNPR. Here I review some evidence that the mechanisms in SNPR are empirically valid. The evidence is provided by the results of running the program on an artificial language sample (see Table 7.2); the details are described in Wolff (1982).

An artificial text with no segmentation markers was prepared from a simple grammar but all instances of two of the (64) sentences generated by the grammar were excluded from the text. When the program was run on this text it successfully retrieved the original grammar despite the fact that the generative range of the grammar was not fully represented in the sample. In the course of building up the grammar it produced many *wrong* generalizations all of which were corrected. Every one of the *correct* generalizations, including those required to predict the missing sentences, were retained as permanent fixtures in the grammar.

In this case, the criterion of *correct* and *incorrect* was the grammar used to create the text. But this use of an artificial grammar to validate the model, although it is justified as an aid to developing the model, is potentially very misleading. The grammar used to create the text is only one of many that could have produced the same text and without some independent criterion there is no guarantee that the one employed is the *best* one. It is a mistake to allow one’s knowledge of English (say) to dictate what is right and wrong when one is dealing with a text which may look superficially like a subset of English but whose *true* structure may be significantly different from English.

A fully satisfactory validation of the generalization and correction mechanism in SNPR is likely to prove difficult. No proper judgment can be made until a model has been developed which can give results with natural language including a satisfactory semantic input. If the mechanism allows wrong generalizations through *(wrong* now in the sense that they are not acceptable to a native speaker of the language) this would be clear evidence against the mechanism. But the model may fail more subtly if it eliminates generalizations that native speakers would in fact accept. Errors like these may be very difficult to detect.

Apart from validating the model against the judgments of native speakers of a natural language it is also necessary to demonstrate that the model does in fact realize the optimization principles on which it is based. This is chiefly a matter of demonstrating that CC increases as learning proceeds and that CC/S_g is maintained at a high level. In order to validate the optimization principles themselves it will be necessary to show that improved performance on measures of optimization correlates with success in discovering satisfactory grammars as judged by native speakers of the language. These are matters for future research.

Explanatory Spin-off

The distinction between explanations considered in this section and those in previous sections is not very clear cut because the prominent facts of language acquisition have been born in mind at all times and they have affected the selection and rejection of hypothesized learning mechanisms. However, what follows was not a primary focus in developing the theory and may reasonably be counted as explanatory bonus, at least in part.

1. The Rate of Acquisition of Words

Children typically produce their first word at about 12 months. In the following 6 months, new words are acquired rather slowly but then the pace quickens to produce a flood of words in the period of 18 months to 3 years. Because many of these new words are object names, this phenomenon is often called the “naming explosion.” The rate at which new words are learned continues to be high throughout most of the rest of childhood; this is not as noticeable as the first burst of activity because it is not quite as dramatic. It is also quite difficult, without special techniques, to assess the size of the person’s vocabulary when it is anything but very small. From early adulthood the rate of acquisition of new words declines progressively into old age.

The picture of vocabulary growth just sketched derives mainly from observations of spoken words but it seems to be similar for receptive vocabulary.

This pattern of vocabulary growth can be explained quite well by the theory. We have supposed that children are, from birth, busy building up bits and pieces of language structure starting with very small primitives. Eventually, one of these pieces reaches word size or near word size and, particularly if it is meaningful, it will be identified by adults as the child’s first word. The reason suggested by the theory for the initial slow growth of vocabulary is that the child is still engaged in constructing the elements from which new words are built. With only a restricted range available, vocabulary growth will be slow. When the range is bigger, the rate of acquisition of new words will increase because relatively little processing is required in the construction of each new word; the rate should remain high for some time.

We may expect an eventual decline because opportunities to observe new words will eventually become rare enough to put a damper on the learning of new words (but see later discussion of this question). There is some evidence that rare words are constructed from a greater variety of constituents than common words and this might also tend to limit the rate of vocabulary growth.

This kind of informal explanation of the way vocabularies grow is not entirely satisfactory because observed patterns of growth depend in a complex way on the processing characteristics of the child and the statistical structure of the language being learned. A better way of matching the theory with empirical data is to construct a working model and see what patterns of vocabulary growth emerge when it operates with natural language. This has been done and the patterns produced correspond quite well with the picture sketched above (Wolff, 1977).

2. The Order of Acquisition of Words and Morphemes

Except for an anomaly to be discussed, there is a clear tendency for children to learn common words before rare ones (Gilhooly & Gilhooly. 1979). Given that in most languages rare words tend to be longer than common words and the variety of rare words is greater than the variety of common words (Zipf, 1935) we would expect children to learn long words later than short ones and we would expect the increases in lengths of words at successive ages to become progressively smaller.

A clear implication of the theory is that common structures should be learned earlier than rare ones. The implication that long words should be acquired later than short ones depends purely on the known relationship between frequency and word length and is nothing to do with the fact that the program builds large structures from small ones: This feature of the program would be entirely compatible with zero or even negative correlation between word length and age of acquisition.

As in the previous section, a working model is needed to see in detail how well expected patterns of acquisition correspond with observed patterns. Program MK10 produces a very good match with available data (Wolff, 1977).

Figure 7.5 shows the relationship between the lengths of words acquired by a young child and the ages at which they were acquired (data from Grant, 1915). Figure 7.6 shows comparable data for program MK10 applied to three different samples of natural language. In both figures, the progressive increase in the lengths of words can be seen and also the progressive decrease in the rate at which lengths increase.

Over the age range covered in Fig. 7.5, the decreasing rate at which word lengths increase is not very obvious. But it is easy to establish that the curve will be less steep at later ages because a linear extrapolation of the curve in Fig. 7.5 would lead one to predict that children would be acquiring absurdly long words in their late childhood and teens.

The anomaly mentioned above is that, in the early stages of language acquisition, function words tend to appear later than content words (McCarthy, 1954)

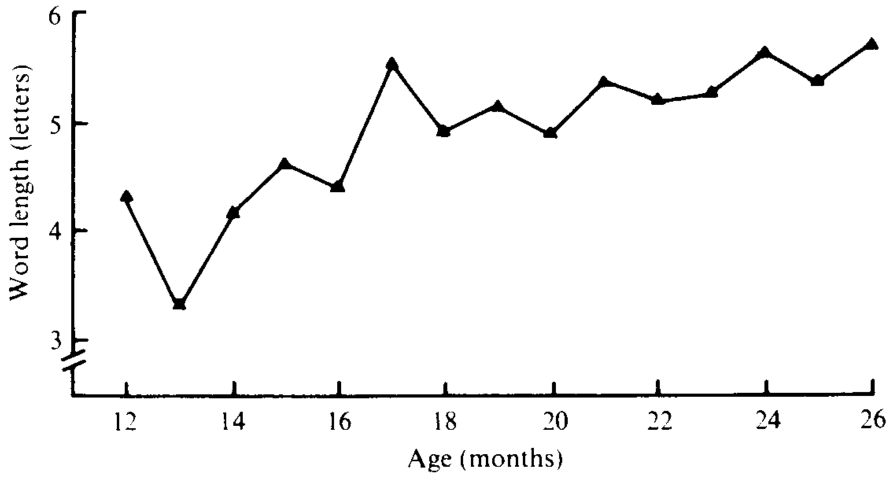


FIG. 7.5. The average lengths of words acquired by one child at different ages (Wolff, 1977; data from Grant, 1915). This figure is reproduced by kind permission of The British Psychological Society.

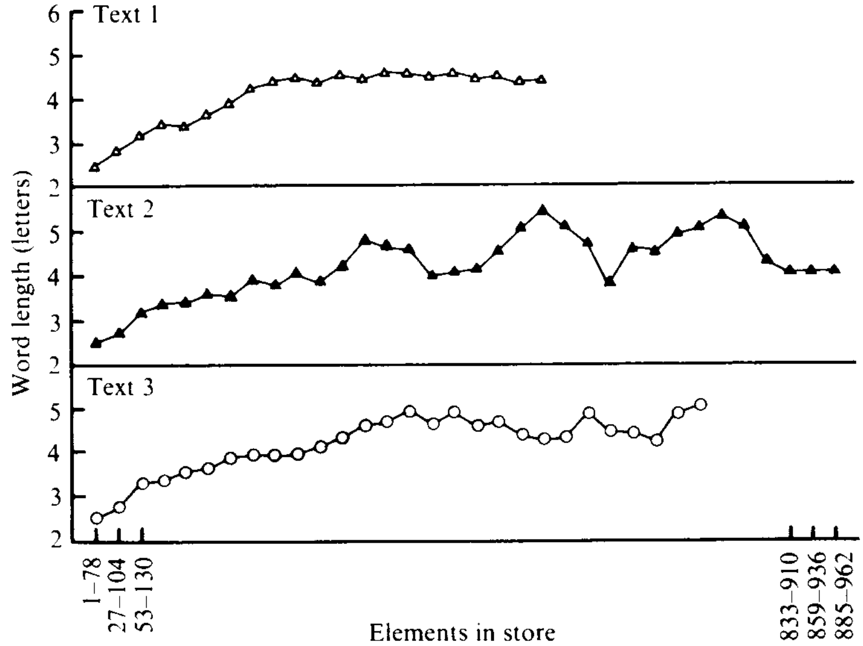


FIG. 7.6. The average lengths of words isolated by program MK10 from three natural language texts at different stages of processing (Wolff, 1977). This figure is reproduced by kind permission of The British Psychological Society.

although they are, typically, amongst the most frequent words in any language (Fries, 1952).

The best available explanation of this exception to the general pattern is that function words are largely meaningless until the larger syntactic patterns in which they function have been built up. A child may know the sound patterns of "and," "the," "into," etc., at an early age but have no cause to use them until they can be fitted into coordination constructions, noun phrases and prepositional phrases respectively. By contrast, words like "table," "Mummy," "more," etc., are quite useful by themselves and can sensibly be used as soon as they are learned. A clear prediction of the theory, then, which would be interesting but difficult to test, is that children do have a knowledge of function words at an age before they start to use them.

Brown's (1973) observation that there is no significant correlation between frequency of use by caretaking adults and order of acquisition of fourteen functional morphemes (e.g., present progressive "-ing," preposition "on," plural "-s" etc.) is completely at odds with the theory presented here. The argument used above (that the later acquisition of certain forms is more apparent than real) cannot be used here because all of the forms considered in this part of Brown's study are alike in that they are all functional morphemes, no one of which can sensibly be used as a meaningful utterance by itself. The criterion of acquisition was, in every case, correct use in 90% of obligatory contexts.

The conflict between the present theory and Brown's results is apparently resolved by the more recent conclusion (Forner, 1979; Moerk, 1980) that Brown's analyses of his data are in fact wrong. When defensible changes and refinements are made in the assumptions that go into the analyses then substantial and significant (negative) correlations are found between frequency of use and order of acquisition of these fourteen morphemes.

3. Brown's (1973) Law of Cumulative Complexity

Perhaps the most interesting general conclusion from Brown's (1973) classic study is that if one structure contains everything that another structure contains *and more* then it will be acquired later than that other structure. Given the variety of current linguistic theories there is some uncertainty about what the *content* of a structure might be but this "Law of Cumulative Complexity" seems to stand up almost regardless of the theoretical framework adopted.

The law is not as trivial as it may at first sight seem although it does correspond with untutored expectations about language acquisition. It is conceivable that children might, in a certain sense, *acquire* structures whose internal organization is, initially, quite unlike the mature form eventually attained. A pattern ABC might be acquired directly or built up as (A(BC)) and then, with the subsequent recognition of AB as a discrete entity, it might be restructured as ((AB)C).

It may at first be thought that Brown's Law follows directly from the way the SNPR model has been designed: the model may be thought to be less an explanation of the law than a restatement of it. It is true that SNPR builds its structures from previously established constituents but there is nothing in the model to prevent a structure being built up initially in one form and then reconstructed in another.

The suggested reason why children (and the model) do not generally do this is that (as with the building up of words) frequency is the guiding heuristic. As a matter of observation (Brown & Hanlon, 1970) complex structures are less frequent than their constituents. According to the theory, they should, therefore, be acquired later.

4. The S-P/Episodic-Semantic Shift

It has been recognized for some time that the way children respond in a word association task changes as they mature. Young children tend to give as their responses words which could follow the stimulus word in a sentence (syntagmatically related words) whereas older children and adults tend to respond with (paradigmatically related) words which can be substituted for the stimulus word in a sentence (see, for example, Entwisle, 1966).

More recently, Petrey (1977) has reexamined Entwisle's data and has argued, persuasively enough, that while the syntagmatic-paradigmatic (S-P) shift remains roughly true, changes in word association responses through childhood may be more accurately characterized as an "episodic-semantic" shift. What this means is that young children tend to give as responses either words which could follow the stimulus word in a sentence or words which signify objects or events which could have been experienced by the child at the same time and place as the object or event signified by the stimulus word, or both of these. A seeming example of a response based on physical contiguity is a child saying "cook" after the stimulus word "add." Petry points out that this superficially bizarre response makes good sense when you see that other responses to "add" include "flour," "milk," "water," "dinner," "cake," etc. *Adding* things is something which young children may well typically first experience in the context of cooking.

Older children and adults tend to give responses which are related to the stimulus word in some way more abstract than mere syntactic or temporal/spatial contiguity. This abstract relationship (sometimes called "semantic") is typically both a paradigmatic (part-of-speech) relationship and a meaning relationship as in "long-short," "wild-tame," "give-take," etc.

If we assume that word association norms at different ages reflect changing organization of stored knowledge then these phenomena make good sense in terms of our theory (see also Kiss, 1973). The theory postulates that children search for recurring clusters of spatially and/or temporally contiguous *events* both in their linguistic input and in their other experience. They also search for groups of elements in which members of the group share one or more temporal or spatial contexts. The disjunctive groups are incorporated in complex elements which represent clusters of similar patterns (see later).

The crucial point here is that the latter kind of search depends on the prior formation of clusters based on contiguity—it cannot get off the ground until there is a big enough set of simple clusters from which to derive common contexts and similarity groupings. If simple contiguity groups correspond to episodic knowledge and disjunctive/complex groupings correspond to semantic knowledge then the delay, just mentioned, in the construction of disjunctive/complex structures provides an explanation of the episodic-semantic shift; we apparently have an answer to Petrey's (1977) question: ". . . by what process can episodic memories of words in context lead to the abstract semantic organisation of mature lexical storage?" (p.70).

There is other evidence supporting the present view of the S-P shift. This shift tends to occur earlier for high frequency stimulus words than for rare ones and it correlates with the variety of syntactic contexts in which a word appears (see Kiss, 1973). The second observation is probably equivalent to the first one given that words with a wide variety of contexts will tend also to be frequent. The late appearance of an S-P shift in rare words may be attributed firstly to the fact that such words are themselves learned late and secondly to the probable fact that they tend to fall in contextual patterns which are less frequent than the most frequent contexts of common words. There are details here that need quantification.

A seeming problem for the account just given is that, as Petrey points out, children are speaking more-or-less correctly by the age of 4 whereas the S-P shift is most dramatic between the ages of 6 and 8. If, as Petrey assumes, correct speech is evidence of a knowledge of part-of-speech categories then it is hard to understand why paradigmatic responding does not appear earlier.

It is plain that children are combining words in a creative way from a very early age and it is tempting to assume, therefore, that all their speech at all stages is produced by combining elements according to rule and guided by a knowledge of permissible substitutes in particular contexts. This is not necessarily so. As least some correct utterances may be produced as essentially direct replicas of utterances previously heard. Both the theory and the observation just mentioned would lead us to expect that young children would produce a relatively high proportion of utterances of this kind. It would be interesting although perhaps methodologically difficult to obtain evidence on this point.

5. Overgeneralizations

The idea that children, in forming linguistic generalizations, might, so to speak, overshoot their target and produce wrong overgeneralizations is not merely a byproduct of the theory. Clear examples of overgeneralization like "hitted," "mouses," etc., are very prominent in young children's speech and they are indeed one of the most salient pieces of evidence that children are abstracting general rules.

The theory not only provides a mechanism for correcting wrong generalizations but it seems to explain a quirk which has been observed in the way these generalizations arise. Children apparently produce irregular plurals and past tense verbs like “geese,” “mice,” “fought,” etc., in their *correct* form initially. *Only later* do they substitute the overgeneralized “gooses,” “mouses,” “fighted,” etc., and then revert eventually to the correct forms again (Slobin, 1971).

This pattern fits the theory well because, as explained in the section on the S-P shift, disjunctive groupings (and the generalizations that derive from them) can only be formed at a stage when there is a range of simple patterns to generalize from. The irregular nouns and verbs will be learned as they are observed and then displaced by generalizations when they are formed. The correction mechanism will restore them later.

6. The S/owing of Language Development in Later Years

The way in which vocabulary growth eventually slows down echoes the more pronounced way in which a child’s learning of grammatical patterns is accomplished largely before the age of 5 and then tails off in later years (see Chomsky, 1969).

A commonsense explanation of these effects would be that the child cannot continue to build up his knowledge of language if he or she has extracted all the available patterns in the data. In order to account for a progressive slowing in language learning rather than a sharp cessation of learning one could refine this view by taking account of the way unlearned structures would become progressively rarer as the data becomes exhausted. The opportunities to observe new structures would become more and more sparse. Notice that this explanation does not depend on the observations that common structures are learned earlier than rare ones although it is entirely compatible with it.

Plausible as this *exhaustion* explanation may appear, it is very probably wrong or at least only partly true.

Although there are many uncertainties and methodological difficulties in estimating the total size of the person’s vocabulary (Ellegard, 1960; Seashore & Eckerson, 1940), it is clear that most people in their lifetimes do not come anywhere near exhausting the word forms in their native language, certainly for a language like English with its exceptionally large vocabulary. While people may reach a stage in their learning where the frequency of any particular unlearned word is very low, the variety of as yet unlearned words is so great that the frequency of this class of words as a whole is relatively high.

There are considerable difficulties in determining the extent of a person’s knowledge of grammatical patterns and there are uncertainties in what should or should not be regarded as a distinct pattern, but it seems reasonably clear that there are many esoteric patterns that people do not generally bother to learn.

"We found 9-year-olds and l0-year-olds who could not, even with prodding, respond with the correct answer: "What should I feed the doll?" [in response to the instruction "Ask L what to feed the doll"]. The question that we wish to raise is whether these children are still in a process of acquisition with respect to this structure and will at some future time be able to interpret it correctly, or whether perhaps they may already have reached what for them constitutes adult competence. We have observed from informal questioning that this structure is a problematic one for many adults, and there are many adult speakers who persist in assigning the wrong subject to the complement verb. This seems to be a structure that is never properly learned by a substantial number of speakers" (Chomsky, 1969, p. 101).

Before we proceed to consider an alternative or supplementary explanation of the slowing down in the acquisition of new language patterns, one other commonsense explanation may be noted. Part of the cause of a slowing up in language learning may be a reduction in processing capacity because of physical deterioration in the brain. Barring disease, such an effect looks unimportant before old age. Even then it would seem to be only a minor factor because of the informal observation that people can pick up new words rapidly at almost any age if they are introduced to a new language.

The explanation suggested by the theory for why language learning slows up in later years has to do with the tradeoffs which are basic in the theory. The two that have been considered so far will serve the argument. Children are supposedly miserly in their use of storage space for long-term storage: New information will only be stored in a long-term form if it adds significantly to the usefulness of the knowledge structures for encoding information. In the early stages of learning, plenty of such patterns are observed and quickly incorporated in the child's long-term knowledge structures. Later on, patterns that are useful enough to warrant long-term storage will be encountered less and less often and acquisition will slow.

We might imagine that this gradual slowing in the growth of linguistic knowledge would have a sharp terminus when the child's database is finally exhausted of all structures that are useful enough to be worth storing. However, this expectation is based on the assumption of a fixed database. Given a continually expanding linguistic corpus and the resulting fluctuations in the observed frequencies of linguistic patterns, we have a second reason for expecting a gradual tailing off in language learning rather than an abrupt end to it.

The essential difference between this *tradeoff* explanation and the *exhaustion* explanation mentioned at the beginning of this section, is that it proposes optimization as a limiting factor rather than the availability of patterns in the data. Given uniform linguistic experiences and given variations in how miserly individuals need to be with storage (this in turn presumable depending in part on the total available storage), the preferred view predicts variations among individuals in how big their mature system will be while the other view does not. The two views differ also in that the preferred view does not require anyone to exhaust the data available to them whereas the other view does.

7. Nonlinguistic Cognitive Structures

As already stated, a working assumption in this project has been that a set of principles may be found to operate in all spheres of knowledge acquisition (which is not to say that differences may not also be found). Nonetheless, most work to date in this project has been done with input data and knowledge structures which are most clearly analogous to syntax. There has been a relatively unsuccessful attempt to develop ideas in the nonlinguistic sphere (Wolff, 1976), this at a stage before several important insights had been achieved. Wolff (1987) argues for a uniform system for encoding syntactic and semantic knowledge but relatively little is said about the latter. Only a little is said here. The topic really warrants a whole paper to itself.

The chief merit of my 1976 paper is to establish a set of target criteria of success for a theory of how classes of objects *(concepts* in this context) may be developed. Briefly, these are:

1. Natural classifications of objects differ from the artificial classes studied by, for example, Bruner, Goodnow, and Austin (1956), in that they are in some sense *salient:* They reflect structures inherent in the world which our concept learning systems can abstract without *explicit teaching.*
2. Our concepts are arranged in hierarchies and heterarchies.
3. There is overlap among conceptual groupings.
4. The boundaries of natural classes are in some sense *fuzzy.*
5. Natural classes are often *polythetic:* No single attribute or group of attributes need be shared by all members of the class. (This together with 3 and 4 above are the chief differences between natural classification systems and those developed by the majority of clustering algorithms.)
6. Attributes of objects carry varying *weights* in the process of recognizing new instances of a class.

The model described in Wolff (1976) was reasonably successful at meeting all the criteria, except the requirement of polythesis. Now it seems that program SNPR, although it was developed primarily as a model of syntax learning, meets all six criteria completely. This is not to say that it is a wholly satisfactory model of concept acquisition—there are other criteria that may be added to these six which it would not be able to model.

The reason that SNPR can be seen as a model of concept learning is that it develops disjunctive classes and it also develops complex elements that can be seen as intensional descriptions of classes of *similar* entities. For example, a complex element with the structure (A|B)X(C|D)Y (where | represents exclusive “OR”) describes the extensional class of entities AXCY, BXCY, AXDY and BXDY. The members of this class are similar, obviously, because they share attributes X and Y and, less obviously, because there is some commonality among them in the attributes A, B, C, and D. Because X and Y are common to all members, this particular class is not polythetic. But SNPR is quite capable of developing intensional descriptions like (A|B)(C|D)(E|F) where the members of the corresponding extensional class (ACE, ACF, BCE, BCF, ADE, ADF, BDE, BDF) have no single attribute in common. This is a truly polythetic category.

The term *attribute* used here need not be confined to conventional perceptual attributes like shapes and colors. It may also cover functional attributes like the fact that a ball can roll (Nelson’s, 1974 example). Contextual properties of concepts—fish are typically found in water, for example—are handled quite straightforwardly by the system because of the way it develops part-whole hierarchies: The concept of fish can be incorporated in a larger element representing fishy environments. Contextual or extrinsic attributes of concepts are arguably equal in importance to conventional intrinsic attributes in establishing the nature of a category.

SNPR also meets the other five criteria. The elements developed by the model are salient in the sense that they express redundancies in the input data and are discovered without explicit teaching. SNPR can develop part-whole hierarchies and it can also develop class-inclusion hierarchies in which overlap between classes can occur. Fuzziness of concepts and differential weighting of attributes can be dealt with by allowing the identification of new entities as belonging to one or other of preestablished categories to be a probabilistic matter (as in my 1976 model).

The relevance of the theory to the realm of nonlinguistic cognitions is underlined by the similarity between the complex elements developed by SNPR and the well-known notions of *schema* (Bartlett, 1932; Bobrow & Norman, 1975), *frame* (Minsky, 1975), and *script* (Schank & Abelson, 1977). Like these theoretical constructs, a complex element in the theory is a generalized pattern which reflects a commonly recurring set of entities, be it a set of cultural expectations (Bartlett) or the things found inside a typical room (Minsky) or the typical pattern of events that occur when you eat a meal in a restaurant (Schank & Abelson). All these notions share the idea that there are *slots* in the framework where alternatives may be inserted, and they share the idea that one of the alternatives may function as a default—the assumed filler for the slot when there is no contrary evidence. In the syntagmatic elements developed by SNPR the disjunctive constituents are equivalent to slots. Since members of each disjunctive set typically vary in their contextual probability, the most probable one may be regarded as a default element.

The foregoing is intended to indicate how a theory largely developed with reference to syntactic phenomena may indeed generalize to semantic phenomena with little if any adjustment, in accordance with the working hypothesis of uniform structure-abstracting principles. The major gap in these ideas is some principled account of the origin and growth of *relational* concepts. This is a matter for future work (but see Wolff, 1987).

8. Nativist Arguments

Three planks of the nativist position have been that a knowledge of language structure must be largely known in advance because the available evidence contained in the language which a child hears is too much obscured by *performance* errors and distortions of various kinds; because the vagaries of individual experience of a given language and individual variations in ability do not square with the way everyone acquires essentially the same grammar; and because language acquisition apparently happens too fast to be explained by learning alone.

> A consideration of the character of the grammar that is acquired, the degenerate quality and narrowly limited extent of the available data, the striking uniformity of the resulting grammars, and their independence of intelligence, motivation, and emotional state, over wide ranges of variation, leave little hope that much of the structure of the language can be learned by an organism initially uninformed as to its general character. (Chomsky, 1965, p. 58)

And later:

> . . . there is surely no reason today for taking seriously a position that attributes a complex human achievement entirely to months (or at most years) of experience, rather than to millions of years of evolution or to principles of neural organization that may be even more deeply grounded in physical law . . . (Chomsky, 1965, p. 59)

That the data available to the child has a “narrowly limited extent” seems to be simply wrong. Anyone who has had any dealings with the recording of what adults say to children or in their presence will know that the quantities of data are enormous.

That the data are very often corrupted in terms of what native speakers with mature knowledge would judge to be correct, is clearly true; attempts to show otherwise seem to be misplaced. The ‘dirty data’ problem is illustrated in Fig. 7.1. A strength of the theory is that it neatly explains how children can learn from such data without being thrown off by errors: Any particular error is, by its nature, rare and so in the search for useful (common) structures, it is discarded along with many other candidate structures. (If an error is not rare it is likely to acquire the status of a dialect or idiolect variation and cease to be regarded as an error.)

In practice, the programs MK10 and SNPR have been found to be quite insensitive to errors (of omission, addition, or substitution) in their data. A good example with respect to omissions is the way SNPR was able to discover a grammar from data containing less than the complete range of terminal strings of that grammar.

It is probably true that the members of any given language community have grammatical systems which are quite similar, one to another. But they are not identical. There are many more-or-less subtle differences between individual systems (Broadbent, 1970). The uniformity of grammatical systems across a wide ability range seen by authors like Lenneberg (1967) may be attributed in part to unsophisticated methods of assessment. With more penetrating techniques like those pioneered by Chomsky (1969) many differences come to light which can otherwise easily be overlooked.

Even though there is wide variation in children’s experience of any given language, we should not be surprised to find quite a lot of similarity between the grammatical systems that they develop. The reason is that, within one language community, children’s experience of language can be varied at the level of particular sentences but quite uniform at the level of the grammatical patterns on which those sentences are modeled and uniform in terms of the words out of which they are constructed. (Without this uniformity it would not be reasonable to say that the children belonged to one single linguistic community.) Abstraction processes like those in SNPR which are guided by constancy (redundancy) are not distracted by idiosyncratic realizations of recurrent patterns.

The third argument, that language development is too fast to be explained by learning mechanisms, need not detain us. The computational power of a child’s brain is clearly huge. The computational demands of current models are quite high but there is no reason to think they are unrealistically high. Given what has already been achieved with only a few hours of a conventional computer’s time, there is every reason to think that with more computing power exercised over months and years this kind of process may discover the full complexity of language structure quite easily.

9. The Word Frequency Effect

One of the most fully documented phenomena in psychology is the observation that a spoken or written word or other perceptual pattern is, in some sense, more easily perceived if it is frequent in the observer’s experience than if it is rare. This effect is rather insensitive to varying modes of testing and to varying measures of perception.

There have been many attempts to explain the effect, all of which necessarily assume that people have a knowledge of the relative frequencies of these perceptual patterns. But none of them suggest any reason *why* people should have this knowledge. Now, in the theory, we have a natural explanation: A knowledge of the relative frequencies of perceptual patterns (linguistic or otherwise) is a by-product of search processes which are, so to speak, *designed* to construct an optimal cognitive system.

Neutral and Disconfirming Evidence

Most of the empirical evidence presented so far is apparently explicable in terms of the theory and most of it provides support for the theory. There are of course many other observations of children's language development about which nothing has been said. This large residual set of observations may be divided into two parts: observations, which are in a sense neutral with respect to the theory, which will be considered briefly, and some that seem to be incompatible with the theory, which are discussed at more length.

An example of the kind of observation which is neutral with respect to the theory would be the particular utterances recorded by Braine (1963). It happened that the child, Andrew, said "all done," "all buttoned," "all clean," among other things, but he might just as well have said "all eaten," "all black," and "all found," etc.; no current theory can explain why Andrew said the particular things he did say. It would require an extraordinarily precise theory of motivation and the like to pin such things down.

There seems at present to be only one class of observations which conflicts directly with the theory. In its current form the theory makes a clear statement that children progress from small structures to larger ones by concatenating contiguous elements. What this means is that an utterance like "hit the ball" may only be built up as *((hit the)ball)* or (more likely) *(hit(the ball))* and children should never produce telegraphic utterances like "hit ball" as they have been observed to do (Brown, 1973). Likewise, at the level of word structure, it should be impossible for a child to learn a *schema* of salient features of a word with interstitial elements missing and then subsequently fill these details in (as claimed by Waterson, 1971). No child should ever say "[byΣ]" at an immature stage in the attainment of the adult word "brush," as Waterson's son was heard to do. In this example there is vowel substitution (which is explicable by the theory in terms of generalization) but the missing 'r' represents a supposedly unbridgeable gap between the beginning and the end of the word.

CONCLUSION

A theory may suffer many ills. It may be a loose sketchy affair which does not allow one to make reliable inferences. It may be trivial in the sense that it does not do much more than redescribe the data it is meant to explain. Or it may be trivial because it is an overgeneral catch-all theory which cannot be falsified. Many theories in psychology are weak because they are applicable to only a narrow range of phenomena, often ones observed in a laboratory setting which may lack "ecological validity" (Neisser, 1976). "Micro theories" like these are usually weak also because they do not suggest any connections with a broader theoretical framework.

These points are made, of course, to introduce the claim that the theory described in this chapter is reasonably free from these defects. Certainly these pitfalls have been born in mind and considerable efforts have been made to avoid them.

As a theory of language development, the theory seems also to fair quite well against the useful criteria proposed by Pinker (1979) for evaluating such theories. It cannot yet meet the most stringent of these criteria: that it should propose mechanisms which are powerful enough to learn a natural language. But it does show promise in this direction.

The second criterion, that the theory should not propose mechanisms that are narrowly adapted to a particular language, seems to be met. It is almost axiomatic that a universal feature of languages is *structure* expressed as *redundancy*. Mechanisms designed to abstract redundancy will thus be quite general in their application. There is always the possibility, of course, that a language may be found containing a type of redundancy not yet brought within the scope of the theory.

Whether or not the kinds of mechanism proposed can learn a language within the same time span as a child cannot be decided with absolute confidence. But, as previously argued, there is no reason for supposing that they cannot. Pinker's fourth criterion—that the theory should not demand information in its database which is not reliably available to children—is certainly met by the theory. Stress has been laid in this chapter on the way only weak assumptions have been made about children's sensory input.

That a theory of language development should have something to say about the phenomena observed when children progress towards a mature knowledge of language is another criterion which the theory meets fairly well. We have seen how patterns of vocabulary growth, the Law of Cumulative Complexity, and other developmental phenomena may be explained by the theory.

The last criterion is that proposed mechanisms should not be wildly inconsistent with what is known about children's cognitive abilities. Again, the theory seems quite satisfactory in this regard. The child is seen as taking repeated samples of data and abstracting linguistic structures from them. No single sample needs to be very large and all potential problems of combinatorial explosion are met by the heuristic devices embodies in the theory. Quite a lot of computation is required but there is no reason to think that it is beyond the scope of the 10^{10} neurones which are available.

Future Work

Probably the most useful first step to take in future development of this theory would be to test and refine the ideas in Wolff (1987) by constructing a new computer model which embodies them. One aim would be to establish more clearly how parsing and production processes may be married to the proposed representational system. But the main goal would be to examine how well the proposed learning principles may operate with this system. At some stage a resolution must be found to the mismatch between the theory and the previously discussed observations on telegraphic speech.

An area that needs closer attention is the application of optimization principles to the acquisition of relational concepts. Also in need of fuller treatment is an examination in quantitative terms of how well different learning mechanisms can serve the optimization goal. Related to this is the need to test whether or not success in optimization correlates with success in discovering *correct* linguistic structures as judged by native speakers of a language.

Most of the empirical predictions of the theory have been tested against observations that are already recorded in the literature. There are however a few predictions from the theory that invite further empirical work. If a suitable testing method could be found, it would be interesting to see whether or not children really do have some kind of knowledge of function words at a stage before they use them. Likewise, it would be interesting to test whether babies do indeed have a developing receptive knowledge of word fragments at stages before they utter their first words as the theory predicts they should.

One other empirical question that deserves attention concerns children's developing knowledge of disjunctive grouping and the extent to which their utterances are constructed from smaller constituents or are direct readouts of stored patterns. The theory predicts, and the evidence from word-association tests confirms, that children's knowledge of distributional equivalences should lag behind their knowledge of acceptable strings of words. It would be useful if a method could be found of establishing, for any given utterance, exactly how it was produced. One might then be able to test the theory's prediction that the building of utterances from smaller constituents should become increasingly important as the child's linguistic knowledge matures.

The ideas discussed in this chapter are not intended to be a new dogma. As with any theory in this area there are too many points of uncertainty to warrant rigid views. The theory does, however, seem to have sufficient merit to serve as a framework for future theoretical and empirical work on linguistic and cognitive development.

SUMMARY

The chapter has provided a broad perspective on an *optimization* theory of language learning and cognitive development, details of which have been considered elsewhere.

The basic idea in the theory is that linguistic and cognitive development is, in large measure, a process of building cognitive structures towards a form which is optimally efficient for the several functions to be served.

The theory assumes among other things that language learning does not depend on overt speaking or gesturing by the child, it does not require any kind of reinforcement or error correction or other intervention by a “teacher” and it does not require graded sequences of language samples. But it may be helped by any of these things.

The theory provides or suggests explanations for a wide range of phenomena. These include the acquisition of segmental structures in language at word, phrase and sentence levels, the acquisition of part-of-speech and other disjunctive categories, generalization of grammatical rules (including recursive generalizations), correction of overgeneralizations (including some observed peculiarities of how overgeneralizations appear in children’s speech), the varying rates at which words and other structures are acquired throughout childhood and beyond, the order of acquisition of words and more complex grammatical structures, the S-P or episodic-semantic shift, the development of semantic / conceptual structures, and some other observations. The theory fits well into a biological framework and is broadly consistent with current thinking about language comprehension and production.

There are some observations which are in conflict with the theory in its present form.

ACKNOWLEDGMENT

I am grateful to Alan Wilkes and Philip Quinlan of the University of Dundee for useful comments on an earlier draft of this chapter.

The chapter was first drafted while I was a lecturer at Dundee. I am grateful to the university for supporting this work and to my colleagues in the Department of Psychology for friendly criticism and stimulating discussion.

Some of the research described in the chapter was supported by a Personal Research Grant to me (HRP8240/ 1(A)) from the British Social Science Research Council.

REFERENCES

Anderson, J. R. (1981). A theory of language acquisition based on general learning principles. *Proceedings of the Seventh International Conference on Artificial Intelligence IJCAI-81,* 97-103.

Bartlett, F. C. (1932). *Remembering: An experimental and social study.* London: Cambridge University Press.

Bobrow, D. G., & Norman. D. A. (1975). Some principles of memory schemata. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding.* New York: Academic Press.

Braine. M. D. S. (1971). On two types of models of the internalization of grammars. In D. I. Slobin (Ed.). *The ontogenesis of grammar.* New York: Academic Press.

Braine, M. D. S. (1963). The ontogeny of English phrase structure: The first phrase. *Language, 39,* 1-13.

Broadbent, D. E. (1970). In defence of empirical psychology. *Bulletin of the British Psychological Society, 23,* 87-96.

Brown, R. (1973). *A first language: The early stages.* Harmondsworth, England: Penguin.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Carr, H. A. (1931). The laws of association. *Psychological Review, 38,* 212-228.

Chomsky, C. (1969). *The acquisition of syntax in children from 5 to 10.* Cambridge, MA: MIT Press.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Coulon, D., & Kayser, D. (1978). Learning criterion and inductive behaviour. *Pattern Recognition, 10*, 19-25.

Deese, J. (1965). *The structure of association in language and thought.* Baltimore, MD: Johns Hopkins University Press.

Ellegard, A. (1960). Estimating vocabulary size. *Word, 16,* 219-244.

Entwisle, D. R. (1966). *Word associations of young children.* Baltimore, MD: Johns Hopkins University Press.

Forner, M. (1979). The mother as LAD: Interaction between order and frequency of parental input and child production. In F. R. Eckman & A. J. Hastings (Eds.), *Studies in first and second language acquisition.* Rowley, MA: Newberry House.

Fries, C. C. *(1952). The structure of English.* London: Longmans.

Fromkin, V. (Ed). (1973). *Speech errors as linguistic evidence.* The Hague: Mouton.

Gammon, E. (1969). Quantitative approximations to the word. *Tijdschrift van het lnstituut voor Toegepaste Linguistiek (Leuven), 5,* 43-61.

Garner, W. R. (1974). *The processing of* information *and structure.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gilhooly, K. J., & Gilhooly, M. L. (1979). *The age of acquisition of words as a factor in verbal tasks.* Final Report to the British Social Science Research Council on Research Grant HR/5318.

Grant, I. R. (1915). A child's vocabulary and its growth. *Pedagogical Seminary, 22,* 183-203.

Harris, Z. S. (1961). *Structural linguistics.* Chicago: University of Chicago Press.

Harris, Z. S. (1970). *Papers in structural and transformational linguistics.* Dordrecht: Reidel.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation, 7,* 1-41.

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worral & G. Curry (Eds.), *The methodology of scientific research programmes.* Philosophical Papers, Vol. I. Cambridge, England: Cambridge University Press.

Langley, P. (1982). Language acquisition through error recovery. *Cognition & Brain Theory, 5,* 211-255.

Lenneberg, E. H. (1967). *Biological foundations of language.* New York: Wiley.

McCarthy, D. (1954). Language development in children. In L. Carmichael (Ed.), *Manual of child psychology.* New York: Wiley.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.

Moerk, E. L. (1980). Relationships between parental frequency and input frequencies and children's language acquisition: A reanalysis of Brown's data. *Journal of Child Language, 7,* 105.

Neisser, U. (1976). *Cognition and reality.* San Francisco: W. H. Freeman.

Nelson, K. (1974). Concept, word and sentence: Inter-relations in acquisition and development. *Psychological Review, 81,* 267-285.

Olivier, D. C. (1968). Stochastic grammars and language acquisition mechanisms. Unpublished doctoral dissertation, Harvard University.

Petrey, S. (1977). Word association and the development of lexical memory. *Cognition, 5,* 57-71.

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

Rosenfeld, A., Huang, H. K., & Schneider, V. B. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information Theory, IT-15(6),* 672-681.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* New York: Wiley.

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabulary. *Journal of Educational Psychology, 31,* 14-38.

Slobin, D. I. (1971). Data for the Symposium. In D. I. Slobin (Ed.), *The ontogenesis of grammar.* New York: Academic Press.

Waterson, N. (1971). Child phonology: A prosodic view. *Journal of Linguistics 7,* 179-211.

Weir, R. (1962). *Language in the crib.* The Hague: Mouton.

Wolff, J. G. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology, 66,* 79-90.

Wolff, J. G. (1976). Frequency, conceptual structure and pattern recognition. *British Journal of Psychology, 67,* 377-390.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology, 68,* 97-106.

Wolff, J. G. (1980). Language acquisition and the discovery of phrase structure. *Language & Speech, 23,* 255-269.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication, 2,* 57-89.

Wolff, J. G. (1987). Cognitive development as optimization. In L. Bolc (Ed.), *Computational models of learning.* Heidelberg: Springer-Verlag.

Zipf, G. K. (1935). *The psycho-biology of language.* Boston: Houghton Mifflin.