CHAPTER 18

On Learning the Past Tenses of English Verbs

D. E. RUMELHART and J. L. McCLELLAND

THE ISSUE

Scholars of language and psycholinguistics have been among the first to stress the importance of rules in describing human behavior. The reason for this is obvious. Many aspects of language can be characterized by rules, and the speakers of natural languages speak the language correctly. Therefore, systems of rules are useful in characterizing what they will and will not say. Though we all make mistakes when we speak, we have a pretty good ear for what is right and what is wrong—and our judgments of correctness—or grammaticality—are generally even easier to characterize by rules than actual utterances.

On the evidence that what we will and won't say and what we will and won't accept can be characterized by rules, it has been argued that, in some sense, we "know" the rules of our language. The sense in which we know them is not the same as the sense in which we know such "rules" as "*i* before e except after *c*," however, since we need not necessarily be able to state the rules explicitly. We know them in a way that allows us to use them to make judgments of grammaticality, it is often said, or to speak and understand, but this knowledge is not in a form or location that permits it to be encoded into a communicable verbal statement. Because of this, this knowledge is said to be *implicit*.

A slight variant of this chapter will appear in B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum (in press).

So far there is considerable agreement. However, the exact characterization of implicit knowledge is a matter of great controversy. One view, which is perhaps extreme but is nevertheless quite clear, holds that the rules of language are stored in explicit form as propositions, and are used by language production, comprehension, and judgment mechanisms. These propositions cannot be described verbally only because they are sequestered in a specialized subsystem which is used in language processing, or because they are written in a special code that only the language processing system can understand. This view we will call the *explicit inaccessible rule* view.

On the explicit inaccessible rule view, language acquisition is thought of as the process of inducing rules. The language mechanisms are thought to include a subsystem—often called the *language acquisition device* (LAD)—whose business it is to discover the rules. A considerable amount of effort has been expended on the attempt to describe how the LAD might operate, and there are a number of different proposals which have been laid out. Generally, though, they share three assumptions:

- The mechanism hypothesizes explicit inaccessible rules.
- Hypotheses are rejected and replaced as they prove inadequate to account for the utterances the learner hears.
- The LAD is presumed to have *innate* knowledge of the possible range of human languages and, therefore, is presumed to consider only hypotheses within the constraints imposed by a set of *linguistic universals*.

The recent book by Pinker (1984) contains a state-of-the-art example of a model based on this approach.

We propose an alternative to explicit inaccessible rules. We suggest that lawful behavior and judgments may be produced by a mechanism in which there is no explicit representation of the rule. Instead, we suggest that the mechanisms that process language and make judgments of grammaticality are constructed in such a way that their performance is characterizable by rules, but that the rules themselves are not written in explicit form anywhere in the mechanism. An illustration of this view, which we owe to Bates (1979), is provided by the honeycomb. The regular structure of the honeycomb arises from the interaction of forces that wax balls exert on each other when compressed. The honeycomb can be described by a rule, but the mechanism which produces it does not contain any statement of this rule.

In our earlier work with the interactive activation model of word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland,

1981, 1982), we noted that lawful behavior emerged from the interactions of a set of word and letter units. Each word unit stood for a particular word and had connections to units for the letters of the word. There were no separate units for common letter clusters and no explicit provision for dealing differently with orthographically regular letter sequences—strings that accorded with the rules of English—as opposed to irregular sequences. Yet the model did behave differently with orthographically regular nonwords than it behaved with words. In fact, the model simulated rather closely a number of results in the word perception literature relating to the finding that subjects perceive letters in orthographically regular letter strings more accurately than they perceive letters in irregular, random letter strings. Thus, the behavior of the model was lawful even though it contained no explicit rules.

It should be said that the pattern of perceptual facilitation shown by the model did not correspond exactly to any system of orthographic rules that we know of. The model produced as much facilitation, for example, for special nonwords like *SLNT*, which are clearly irregular, as it did for matched regular nonwords like *SLET*. Thus, it is not correct to say that the model exactly mimicked the behavior we would expect to emerge from a system which makes use of explicit orthographic rules. However, neither do human subjects. Just like the model, they showed equal facilitation for vowelless strings like *SLNT* as for regular nonwords like *SLET*. Thus, human perceptual performance seems, in this case at least, to be characterized only approximately by rules.

Some people have been tempted to argue that the behavior of the model shows that we can do without linguistic rules. We prefer, however, to put the matter in a slightly different light. There is no denying that rules still provide a fairly close characterization of the performance of our subjects. And we have no doubt that rules are even more useful in characterizations of sentence production, comprehension, and grammaticality judgments. We would only suggest that parallel distributed processing models may provide a mechanism sufficient to capture lawful behavior, without requiring the postulation of explicit but inaccessible rules. Put succinctly, our claim is that PDP models provide an alternative to the explicit but inaccessible rules account of implicit knowledge of rules.

We can anticipate two kinds of arguments against this kind of claim. The first kind would claim that although certain types of rule-guided behavior might emerge from PDP models, the models simply lack the computational power needed to carry out certain types of operations which can be easily handled by a system using explicit rules. We believe that this argument is simply mistaken. We discuss the issue of computational power of PDP models in Chapter 4. Some applications of PDP models to sentence processing are described in Chapter 19.

The second kind of argument would be that the details of language behavior, and, indeed, the details of the language acquisition process, would provide unequivocal evidence in favor of a system of explicit rules.

It is this latter kind of argument we wish to address in the present chapter. We have selected a phenomenon that is often thought of as demonstrating the acquisition of a linguistic rule. And we have developed a parallel distributed processing model that learns in a natural way to behave in accordance with the rule, mimicking the general trends seen in the acquisition data.

THE PHENOMENON

The phenomenon we wish to account for is actually a sequence of three stages in the acquisition of the use of past tense by children learning English as their native tongue. Descriptions of development of the use of the past tense may be found in Brown (1973), Ervin (1964), and Kuczaj (1977).

In Stage 1, children use only a small number of verbs in the past tense. Such verbs tend to be very high-frequency words, and the majority of these are irregular. At this stage, children tend to get the past tenses of these words correct if they use the past tense at all. For example, a child's lexicon of past-tense words at this stage might consist of *came, got, gave, looked, needed, took*, and *went*. Of these seven verbs, only two are regular—the other five are generally idiosyncratic examples of irregular verbs. In this stage, there is no evidence of the use of the rule—it appears that children simply know a small number of separate items.

In Stage 2, evidence of implicit knowledge of a linguistic rule emerges. At this stage, children use a much larger number of verbs in the past tense. These verbs include a few more irregular items, but it turns out that the majority of the words at this stage are examples of the *regular* past tense in English. Some examples are *wiped* and *pulled*.

The evidence that the Stage 2 child actually has a linguistic rule comes not from the mere fact that he or she knows a number of regular forms. There are two additional and crucial facts:

- The child can now generate a past tense for an invented word. For example, Berko (1958) has shown that if children can be convinced to use *rick* to describe an action, they will tend to say *ricked* when the occasion arises to use the word in the past tense.

- Children now *incorrectly* supply regular past-tense endings for words which they used correctly in Stage 1. These errors may involve either adding *ed* to the root as in *comed* /kˆmd/, or adding *ed* to the irregular past tense form as in *camed* /kAmd/[1] (Ervin, 1964; Kuczaj, 1977).

Such findings have been taken as fairly strong support for the assertion that the child at this stage has acquired the past-tense "rule." To quote Berko (1958):

> If a child knows that the plural of *witch* is *witches*, he may simply have memorized the plural form. If, however, he tells us that the plural of *gutch* is *gutches*, we have evidence that he actually knows, albeit unconsciously, one of those rules which the descriptive linguist, too, would set forth in his grammar. (p. 151)

In Stage 3, the regular and irregular forms coexist. That is, children have regained the use of the correct irregular forms of the past tense, while they continue to apply the regular form to new words they learn. Regularizations persist into adulthood—in fact, there is a class of words for which either a regular or an irregular version are both considered acceptable—but for the commonest irregulars such as those the child acquired first, they tend to be rather rare. At this stage there are some clusters of exceptions to the basic, regular past-tense pattern of English. Each cluster includes a number of words that undergo identical changes from the present to the past tense. For example, there is a *ing/ang* cluster, an *ing/ung* cluster, an *eet/it* cluster, etc. There is also a group of words ending in /d/ or /t/ for which the present and past are identical.

Table 1 summarizes the major characteristics of the three stages.

Variability and Gradualness

The characterization of past-tense acquisition as a sequence of three stages is somewhat misleading. It may suggest that the stages are clearly demarcated and that performance in each stage is sharply distinguished from performance in other stages.

---

[1] The notation of phonemes used in this chapter is somewhat nonstandard. It is derived from the computer-readable dictionary containing phonetic transcriptions of the verbs used in the simulations. A key is given in Table 5.

TABLE 1

CHARACTERISTICS OF THE THREE STAGES OF PAST TENSE ACQUISITION

Verb Type	Stage 1	Stage 2	Stage 3
Early Verbs	Correct	Regularized	Correct
Regular	—	Correct	Correct
Other Irregular	—	Regularized	Correct or Regularized
Novel	—	Regularized	Regularized

In fact, the acquisition process is quite gradual. Little detailed data exists on the transition from Stage 1 to Stage 2, but the transition from Stage 2 to Stage 3 is quite protracted and extends over several years (Kuczaj, 1977). Further, performance in Stage 2 is extremely variable. Correct use of irregular forms is never completely absent, and the same child may be observed to use the correct past of an irregular, the base+ed form, and the past+ed form, within the same conversation.

Other Facts About Past-Tense Acquisition

Beyond these points, there is now considerable data on the detailed types of errors children make throughout the acquisition process, both from Kuczaj (1977) and more recently from Bybee and Slobin (1982). We will consider aspects of these findings in more detail below. For now, we mention one intriguing fact: According to Kuczaj (1977), there is an interesting difference in the errors children make to irregular verbs at different points in Stage 2. Early on, regularizations are typically of the base+ed form, like *goed*; later on, there is a large increase in the frequency of past+ed errors, such as *wented*.

THE MODEL

The goal of our simulation of the acquisition of past tense was to simulate the three-stage performance summarized in Table 1, and to see whether we could capture other aspects of acquisition. In particular, we wanted to show that the kind of gradual change characteristic of normal acquisition was also a characteristic of our distributed model, and we wanted to see whether the model would capture detailed aspects

of the phenomenon, such as the change in error type in later phases of development and the change in differences in error patterns observed for different types of words.

We were not prepared to produce a full-blown language processor that would learn the past tense from full sentences heard in everyday experience. Rather, we have explored a very simple past-tense learning environment designed to capture the essential characteristics necessary to produce the three stages of acquisition. In this environment, the model is presented, as learning experiences, with pairs of inputs—one capturing the phonological structure of the root form of a word and the other capturing the phonological structure of the correct past-tense version of that word. The behavior of the model can be tested by giving it just the root form of a word and examining what it generates as its "current guess" of the corresponding past-tense form.

Structure of the Model

The basic structure of the model is illustrated in Figure 1. The model consists of two basic parts: (a) a simple *pattern associator* network similar to those studied by Kohonen (1977; 1984; see Chapter 2) which learns the relationships between the base form and the past-tense

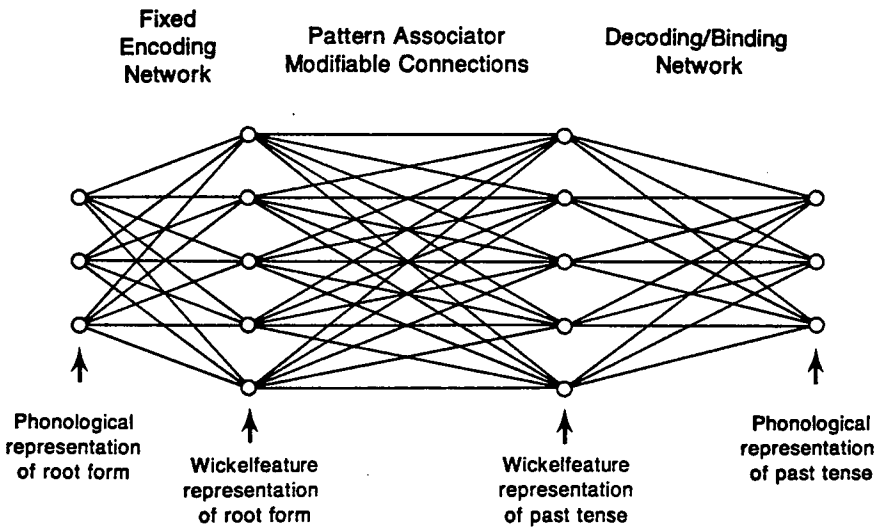


FIGURE 1. The basic structure of the model.

form, and (b) a decoding network that converts a featural representation of the past-tense form into a phonological representation. All learning occurs in the pattern associator; the decoding network is simply a mechanism for converting a featural representation which may be a near miss to any phonological pattern into a legitimate phonological representation. Our primary focus here is on the pattern associator. We discuss the details of the decoding network in the Appendix.

Units. The pattern associator contains two pools of units. One pool, called the input pool, is used to represent the input pattern corresponding to the root form of the verb to be learned. The other pool, called the output pool, is used to represent the output pattern generated by the model as its current guess as to the past tense corresponding to the root form represented in the inputs.

Each unit stands for a particular feature of the input or output string. The particular features we used are important to the behavior of the model, so they are described in a separate section below.

Connections. The pattern associator contains a modifiable connection linking each input unit to each output unit. Initially, these connections are all set to 0 so that there is no influence of the input units on the output units. Learning, as in other PDP models described in this book, involves modification of the strengths of these interconnections, as described below.

Operation of the Model

On test trials, the simulation is given a phoneme string corresponding to the root of a word. It then performs the following actions. First, it encodes the root string as a pattern of activation over the input units. The encoding scheme used is described below. Node activations are discrete in this model, so the activation values of all the units that should be on to represent this word are set to 1, and all the others are set to 0. Then, for each output unit, the model computes the net input to it from all of the weighted connections from the input units. The net input is simply the sum over all input units of the input unit activation times the corresponding weight. Thus, algebraically, the net input to output unit i is

$$net_i = \sum_j a_j w_{ij}$$

where a_j represents the activation of input unit j, and w_{ij} represents the weight from unit j to unit i.

Each unit has a threshold, θ, which is adjusted by the learning procedure that we will describe in a moment. The probability that the unit is turned on depends on the amount the net input exceeds the threshold. The *logistic* probability function is used here as in the Boltzmann machine (Chapter 7) and in harmony theory (Chapter 6) to determine whether the unit should be turned on. The probability is given by

$$p\ (a_i = 1) = \frac{1}{1 + e^{-(net_i - \theta_i)/T}} \tag{1}$$

where T represents the temperature of the system. The logistic function is shown in Figure 2. The use of this probabilistic response rule allows the system to produce different responses on different occasions with the same network. It also causes the system to learn more slowly so the effect of regular verbs on the irregulars continues over a much longer period of time. As discussed in Chapter 2, the temperature, T, can be manipulated so that at very high temperatures the response of the units is highly variable; with lower values of T, the units behave more like *linear threshold units.*

Since the pattern associator built into the model is a one-layer net with no feedback connections and no connections from one input unit to another or from one output unit to another, iterative computation is of no benefit. Therefore, the processing of an input pattern is a simple matter of first calculating the net input to each output unit and then

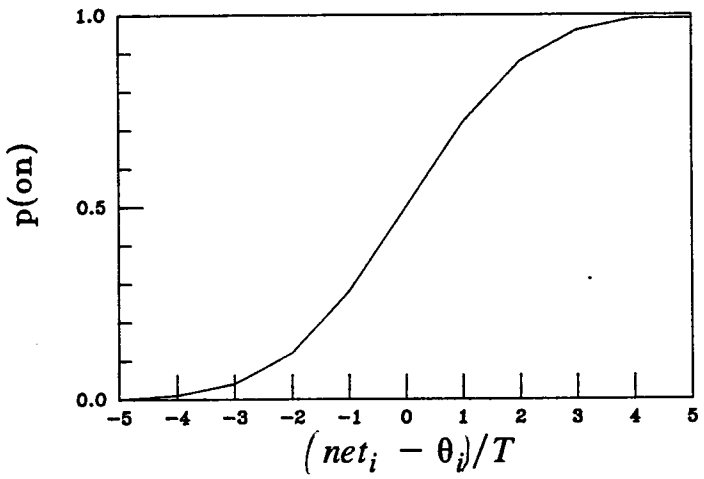


FIGURE 2. The logistic function used to calculate probability of activation. The x-axis shows values of $net_i - \theta_i/T$, and the y-axis indicates the corresponding probability that unit i will be activated.

setting its activation probabilistically on the basis of the logistic equation given above. The temperature T only enters in setting the variability of the output units; a fixed value of T was used throughout the simulations.

To determine how well the model did at producing the correct output, we simply compare the pattern of output Wickelphone activations to the pattern that the correct response would have generated. To do this, we first translate the correct response into a target pattern of activation for the output units, based on the same encoding scheme used for the input units. We then compare the obtained pattern with the target pattern on a unit-by-unit basis. If the output perfectly reproduces the target, then there should be a 1 in the output pattern wherever there is a 1 in the target. Such cases are called *hits*, following the conventions of signal detection theory (Green & Swets, 1966). There should also be a 0 in the output whenever there is a 0 in the target. Such cases are called *correct rejections*. Cases in which there are 1s in the output but not in the target are called *false alarms*, and cases in which there are 0s in the output that should be present in the input are called *misses*. A variety of measures of performance can be computed. We can measure the percentage of output units that match the correct past tense, or we can compare the output to the pattern for any other response alternative we might care to evaluate. This allows us to look at the output of the system independently of the decoding network. We can also employ the decoding network and have the system synthesize a phonological string. We can measure the performance of the system either at the featural level or at the level of strings of phonemes. We shall employ both of these mechanisms in the evaluation of different aspects of the overall model.

Learning

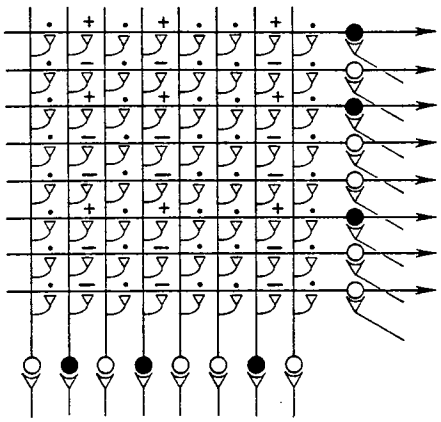
On a learning trial, the model is presented with both the root form of the verb and the target. As on a test trial, the pattern associator network computes the output it would generate from the input. Then, for each output unit, the model compares its answer with the target. Connection strengths are adjusted using the classic *perceptron convergence procedure* (Rosenblatt, 1962). The perceptron convergence procedure is simply a discrete variant of the delta rule presented in Chapter 2 and discussed in many places in this book. The exact procedure is as follows: We can think of the target as supplying a teaching input to each output unit, telling it what value it ought to have. When the actual output matches the target output, the model is doing the right thing

and so none of the weights on the lines coming into the unit are adjusted. When the computed output is 0 and the target says it should be 1, we want to increase the probability that the unit will be active the next time the same input pattern is presented. To do this, we increase the weights from all of the input units that are active by a small amount η. At the same time, the threshold is also reduced by η. When the computed output is 1 and the target says it should be 0, we want to decrease the probability that the unit will be active the next time the same input pattern is presented. To do this, the weights from all of the input units that are active are reduced by η, and the threshold is increased by η. In all of our simulations, the value of η is simply set to 1. Thus, each change in a weight is a unit change, either up or down. For nonstochastic units, it is well known that the perceptron convergence procedure will find a set of weights that will allow the model to get each output unit correct, provided that such a set of weights exists. For the stochastic case, it is possible for the learning procedure to find a set of weights that will make the probability of error as low as desired. Such a set of weights exists if a set of weights exists that will always get the right answer for nonstochastic units.

Learning Regular and Exceptional Patterns in a Pattern Associator

In this section, we present an illustration of the behavior of a simple pattern associator model. The model is a scaled-down version of the main simulation described in the next section. We describe the scaled-down version first because in this model it is possible to actually examine the matrix of connection weights, and from this to see clearly how the model works and why it produces the basic three-stage learning phenomenon characteristic of acquisition of the past tense. Various aspects of pattern associator networks are described in a number of places in this book (Chapters 1, 2, 8, 9, 11, and 12, in particular) and elsewhere (J. A. Anderson, 1973, 1977; J. A. Anderson, Silverstein, Ritz, & Jones, 1977; Kohonen, 1977, 1984). Here we focus our attention on their application to the representation of rules for mapping one set of patterns into another.

For the illustration model, we use a simple network of eight input and eight output units and a set of connections from each input unit to each output unit. The network is illustrated in Figure 3. The network is shown with a set of connections sufficient for associating the pattern of activation illustrated on the input units with the pattern of activation illustrated on the output units. (Active units are darkened; positive



+15	+15					+15	
−15	−15					−15	
+15	+15					+15	
−15	−15					−15	
−15	−15					−15	
+15	+15					+15	
−15	−15					−15	
−15	−15					−15	

FIGURE 3. Simple network used in illustrating basic properties of pattern associator networks; excitatory and inhibitory connections needed to allow the active input pattern to produce the illustrated output pattern are indicated with + and −. Next to the network, the matrix of weights indicating the strengths of the connections from each input unit to each output unit. Input units are indexed by the column they appear in; output units are indexed by row.

and negative connections are indicated by numbers written on each connection). Next to the network is the matrix of connections abstracted from the actual network itself, with numerical values assigned to the positive and negative connections. Note that each weight is located in the matrix at the point where it occurred in the actual network diagram. Thus, the entry in the ith row of the jth column indicates the connection w_{ij} from the jth input unit to the ith output unit.

Using this diagram, it is easy to compute the net inputs that will arise on the output units when an input pattern is presented. For each output unit, one simply scans across its rows and adds up all the weights found in columns associated with active input units. (This is exactly what the simulation program does!) The reader can verify that when the input pattern illustrated in the left-hand panel is presented, each output unit that should be on in the output pattern receives a net input of +45; each output unit that should be off receives a net input of −45.[2] Plugging these values into Equation 1, using a temperature

[2] In the examples we will be considering in this section, the thresholds of the units are fixed at 0. Threshold terms add an extra degree of freedom for each output unit and allow the unit to come on in the absence of input, but they are otherwise inessential to the operation of the model. Computationally, they are equivalent to an adjustable weight to an extra input unit that is always on.

of 15,[3] we can compute that each output unit will take on the correct value about 95% of the time. The reader can check this in Figure 2; when the net input is +45, the exponent in the denominator of the logistic function is 3, and when the net input is −45, the exponent is −3. These correspond to activation probabilities of about .95 and .05, respectively.

One of the basic properties of the pattern associator is that it can store the connections appropriate for mapping a number of different input patterns to a number of different output patterns. The perceptron convergence procedure can accommodate a number of arbitrary associations between input patterns and output patterns, as long as the input patterns form a linearly independent set (see Chapters 9 and 11). Table 2 illustrates this aspect of the model. The first two cells of the table show the connections that the model learns when it is trained on each of the two indicated associations separately. The third cell shows connections learned by the model when it is trained on both patterns in alternation, first seeing one and then seeing the other of the two. Again, the reader can verify that if either input pattern is presented to a network with this set of connections, the correct corresponding output pattern is reconstructed with high probability; each output unit that should be on gets a net input of at least +45, and each output unit that should be off gets a net input below −45.

The restriction of networks such as this to linearly independent sets of patterns is a severe one since there are only N linearly independent patterns of length N. That means that we could store at most eight unrelated associations in the network and maintain accurate performance. However, if the patterns all conform to a general rule, the capacity of the network can be greatly enhanced. For example, the set of connections shown in Table 2D is capable of processing all of the patterns defined by what we call the *rule of 78*. The rule is described in Table 3. There are 18 different input/output pattern pairs corresponding to this rule, but they present no difficulty to the network. Through repeated presentations of examples of the rule, the perceptron convergence procedure learned the set of weights shown in cell D of Table 2. Again, the reader can verify that it works for any legal association fitting the rule of 78. (Note that for this example, the "regular" pairing

[3] For the actual simulations of verb learning, we used a value of T equal to 200. This means that for a fixed value of the weight on an input line, the effect of that line being active on the unit's probability of firing is much lower than it is in these illustrations. This is balanced by the fact that in the verb learning simulations, a much larger number of inputs contribute to the activation of each output unit. Responsibility for turning a unit on is simply more distributed when larger input patterns are used.

TABLE 2

WEIGHTS IN THE 8-UNIT NETWORK
AFTER VARIOUS LEARNING EXPERIENCES

A. Weights acquired in learning (2 4 7) → (1 4 6)

.	15	.	15	.	.	15	.
.	-16	.	-16	.	.	-16	.
.	-17	.	-17	.	.	-17	.
.	16	.	16	.	.	16	.
.	-16	.	-16	.	.	-16	.
.	17	.	17	.	.	17	.
.	-16	.	-16	.	.	-16	.
.	-17	.	-17	.	.	-17	.

B. Weights acquired in learning (3 4 6) → (3 6 7)

.	.	.	-16	-16	.	-16	.
.	.	.	-17	-17	.	-17	.
.	.	.	17	17	.	17	.
.	.	.	-16	-16	.	-16	.
.	.	.	-17	-17	.	-17	.
.	.	.	16	16	.	16	.
.	.	.	17	17	.	17	.
.	.	.	-17	-17	.	-17	.

C. Weights acquired in learning A and B together

.	24	-24	.	.	-24	24	.
.	-13	-13	-26	.	-13	-13	.
.	-23	24	1	.	24	-23	.
.	24	-25	-1	.	-25	24	.
.	-13	-13	-26	.	-13	-13	.
.	13	13	26	.	13	13	.
.	-25	24	-1	.	24	-25	.
.	-12	-13	-25	.	-13	-12	.

D. Weights acquired in learning the rule of 78

61	-37	-37	-5	-5	-3	-6	-7
-35	60	-38	-4	-6	-3	-5	-8
-39	-35	61	-4	-5	-4	-7	-6
-6	-4	-5	59	-37	-37	-8	-7
-5	-5	-4	-36	60	-38	-7	-7
-5	-4	-6	-37	-38	60	-8	-7
.	1	.	1	.	.	-50	51
.	.	-1	-2	1	.	49	-50

TABLE 3

THE RULE OF 78

Input patterns consist of one active unit from each of the following sets:	(1 2 3) (4 5 6) (7 8)
The output pattern paired with a given input pattern consists of:	The same unit from (1 2 3) The same unit from (4 5 6) The other unit from (7 8)
Examples:	2 4 7 → 2 4 8 1 6 8 → 1 6 7 3 5 7 → 3 5 8
An exception:	1 4 7 → 1 4 7

of (1 4 7) with (1 4 8) was used rather than the exceptional mapping illustrated in Table 3).

We have, then, observed an important property of the pattern associator: If there is some structure to a set of patterns, the network may be able to learn to respond appropriately to all of the members of the set. This is true, even though the input vectors most certainly do not form a linearly independent set. The model works anyway because the response that the model should make to some of the patterns can be predicted from the responses that it should make to others of the patterns.

Now let's consider a case more like the situation a young child faces in learning the past tenses of English verbs. Here, there is a regular pattern, similar to the rule of 78. In addition, however, there are exceptions. Among the first words the child learns are many exceptions, but as the child learns more and more verbs, the proportion that are regular increases steadily. For an adult, the vast majority of verbs are regular.

To examine what would happen in a pattern associator in this kind of a situation, we first presented the illustrative 8-unit model with two pattern pairs. One of these was a regular example of the 78 rule [(2 5 8)→(2 5 7)]. The other was an exception to the rule [(1 4 7)→(1 4 7)]. The simulation saw both pairs 20 times, and connection strengths were adjusted after each presentation. The resulting set of connections is shown in cell A of Table 4. This number of learning trials is not enough to lead to perfect performance; but after this much experience, the model tends to get the right answer for each output unit close to 90 percent of the time. At this point, the fact that one of the patterns is an example of a general rule and the other is an exception to that rule is irrelevant to the model. It learns a set of connections that can accommodate these two patterns, but it cannot generalize to new instances of the rule.

This situation, we suggest, characterizes the situation that the language learner faces early on in learning the past tense. The child knows, at this point, only a few high-frequency verbs, and these tend, by and large, to be irregular, as we shall see below. Thus each is treated by the network as a separate association, and very little generalization is possible.

But as the child learns more and more verbs, the proportion of regular verbs increases. This changes the situation for the learning model. Now the model is faced with a number of examples, all of which follow the rule, as well as a smattering of irregular forms. This new situation changes the experience of the network, and thus the pattern of interconnections it contains. Because of the predominance of the regular

TABLE 4

REPRESENTING EXCEPTIONS: WEIGHTS IN THE 8-UNIT NETWORK

A. After 20 exposures to (1 4 7)→(1 4 7), (2 5 8)→(2 5 7)

12	-12	.	12	-12	.	12	-12
-11	13	.	-11	13	.	-11	13
-11	-11	.	-11	-11	.	-11	-11
12	-12	.	12	-12	.	12	-12
-11	11	.	-11	11	.	-11	11
-11	-12	.	-11	-12	.	-11	-12
12	11	.	12	11	.	12	11
-11	-13	.	-11	-13	.	-11	-13

B. After 10 more exposures to all 18 associations

44	-34	-26	-2	-10	-4	-8	-8
-32	46	-27	-11	2	-4	-9	-4
-30	-24	43	-5	-5	-1	-2	-9
-1	-7	-7	45	-34	-26	-4	-11
-8	-3	-3	-31	44	-27	-7	-7
-6	-8	-3	-31	-28	42	-7	-10
11	-2	-6	11	-2	-6	-35	38
-9	-4	7	-13	1	6	36	-42

C. After 30 more exposures to all 18 associations

61	-38	-38	-6	-5	-4	-6	-9
-38	62	-39	-6	-5	-4	-8	-7
-37	-38	62	-5	-5	-3	-7	-6
-4	-6	-6	62	-40	-38	-8	-8
-5	-5	-4	-38	62	-38	-7	-7
-6	-4	-5	-38	-39	62	-8	-7
20	-5	-4	22	-5	-6	-50	61
-19	8	5	-18	5	7	54	-60

D. After a total of 500 exposures to all 18 associations

64	-39	-39	-5	-4	-5	-7	-7
-39	63	-39	-5	-5	-5	-7	-8
-39	-40	64	-5	-5	-5	-8	-7
-5	-5	-5	64	-40	-39	-8	-7
-5	-5	-5	-39	63	-39	-7	-8
-5	-5	-5	-39	-39	63	-8	-7
71	-28	-29	70	-28	-28	-92	106
-70	27	28	-70	27	28	91	-106

form in the input, the network learns the regular pattern, temporarily "overregularizing" exceptions that it may have previously learned.

Our illustration takes this situation to an extreme, perhaps, to illustrate the point. For the second stage of learning, we present the model with the entire set of eighteen input patterns consisting of one active unit from (1 2 3), one from (4 5 6), and one from (7 8). All of these patterns are regular except the one exception already used in the first stage of training.

At the end of 10 exposures to the full set of 18 patterns, the model has learned a set of connection strengths that predominantly captures the "regular pattern." At this point, its response to the exceptional pattern is *worse* than it was before the beginning of Phase 2; rather than getting the right output for Units 7 and 8, the network is now *regularizing* it.

The reason for this behavior is very simple. All that is happening is that the model is continually being bombarded with learning experiences directing it to learn the rule of 78. On only one learning trial out of 18 is it exposed to an exception to this rule.

In this example, the deck has been stacked very strongly against the exception. For several learning cycles, it is in fact quite difficult to tell from the connections that the model is being exposed to an exception mixed in with the regular pattern. At the end of 10 cycles, we can see that the model is building up extra excitatory connections from input Units 1 and 4 to output Unit 7 and extra inhibitory strength from Units 1 and 4 to Unit 8, but these are not strong enough to make the model get the right answer for output Units 7 and 8 when the (1 4 7) input pattern is shown. Even after 40 trials (panel C of Table 4), the model still gets the wrong answer on Units 7 and 8 for the (1 4 7) pattern more than half the time. (The reader can still be checking these assertions by computing the net input to each output unit that would result from presenting the (1 4 7) pattern.)

It is only after the model has reached the stage where it is making very few mistakes on the 17 regular patterns that it begins to accommodate to the exception. This amounts to making the connection from Units 1 and 4 to output Unit 7 strongly excitatory and making the connections from these units to output Unit 8 strongly inhibitory. The model must also make several adjustments to other connections so that the adjustments just mentioned do not cause errors on regular patterns similar to the exceptions, such as (1 5 7), (2 4 7), etc. Finally, in panel D, after a total of 500 cycles through the full set of 18 patterns, the weights are sufficient to get the right answer nearly all of the time. Further improvement would be very gradual since the network makes errors so infrequently at this stage that there is very little opportunity for change.

It is interesting to consider for a moment how an association is represented in a model like this. We might be tempted to think of the representation of an association as the difference between the set of connection strengths needed to represent a set of associations that includes the association and the set of strengths needed to represent the same set excluding the association of interest. Using this definition, we see that the representation of a particular association is far from invariant. What this means is that learning that occurs in one situation (e.g., in which there is a small set of unrelated associations) does not necessarily transfer to a new situation (e.g., in which there are a number of regular associations). This is essentially why the early learning our illustrative model exhibits of the (1 4 7) → (1 4 7) association in the context of just one other association can no longer support correct performance when the larger ensemble of regular patterns is introduced.

Obviously, the example we have considered in this section is highly simplified. However, it illustrates several basic facts about pattern associators. One is that they tend to exploit regularity that exists in the mapping from one set of patterns to another. Indeed, this is one of the

main advantages of the use of distributed representations. Second, they allow exceptions and regular patterns to coexist in the same network. Third, if there is a predominant regularity in a set of patterns, this can swamp exceptional patterns until the set of connections has been acquired that captures the predominant regularity. Then further, gradual tuning can occur that adjusts these connections to accommodate both the regular patterns and the exception. These basic properties of the pattern associator model lie at the heart of the three-stage acquisition process, and account for the gradualness of the transition from Stage 2 to Stage 3.

Featural Representations of Phonological Patterns

The preceding section describes basic aspects of the behavior of the pattern associator model and captures fairly well what happens when a pattern associator is applied to the processing of English verbs, following a training schedule similar to the one we have just considered for the acquisition of the rule of 78. There is one caveat, however: The input and target patterns—the base forms of the verbs and the correct past tenses of these verbs—must be represented in the model in such a way that the features provide a convenient basis for capturing the regularities embodied in the past-tense forms of English verbs. Basically, there were two considerations:

- We needed a representation that permitted a differentiation of all of the root forms of English and their past tenses.
- We wanted a representation that would provide a natural basis for generalizations to emerge about what aspects of a present tense correspond to what aspects of the past tense.

A scheme which meets the first criterion, but not the second, is the scheme proposed by Wickelgren (1969). He suggested that words should be represented as sequences of context-sensitive phoneme units, which represent each phone in a word as a triple, consisting of the phone itself, its predecessor, and its successor. We call these triples *Wickelphones*. Notationally, we write each Wickelphone as a triple of phonemes, consisting of the central phoneme, subscripted on the left by its predecessor and on the right by its successor. A phoneme occurring at the beginning of a word is preceded by a special symbol (#) standing for the word boundary; likewise, a phoneme occurring at the

end of a word is followed by #. The word /kat/, for example, would be represented as ${}_{\#}\mathbf{k}_{\mathbf{a}}$, ${}_{\mathbf{k}}\mathbf{a}_{\mathbf{t}}$, and ${}_{\mathbf{a}}\mathbf{t}_{\#}$. Though the Wickelphones in a word are not strictly position specific, it turns out that (a) few words contain more than one occurrence of any given Wickelphone, and (b) there are no two words we know of that consist of the same sequence of Wickelphones. For example, /slit/ and /silt/ contain no Wickelphones in common.

One nice property of Wickelphones is that they capture enough of the context in which a phoneme occurs to provide a sufficient basis for differentiating between the different cases of the past-tense rule and for characterizing the contextual variables that determine the subregularities among the irregular past-tense verbs. For example, the word-final phoneme that determines whether we should add /d/, /t/ or /ˆd/ in forming the regular past. And it is the sequence ${}_{\mathbf{i}}\mathbf{N}_{\#}$ which is transformed to ${}_{\mathbf{a}}\mathbf{N}_{\#}$ in the *ing* → *ang* pattern found in words like *sing*.

The trouble with the Wickelphone solution is that there are too many of them, and they are too specific. Assuming that we distinguish 35 different phonemes, the number of Wickelphones would be 35^3, or 42,875, not even counting the Wickelphones containing word boundaries. And, if we postulate one input unit and one output unit in our model for each Wickelphone, we require rather a large connection matrix (4.3×10^4 squared, or about 2×10^9) to represent all their possible connections.

Obviously, a more compact representation is required. This can be obtained by representing each Wickelphone as a distributed pattern of activation over a set of feature detectors. The basic idea is that we represent each phoneme, not by a single Wickelphone, but by a pattern of what we call *Wickelfeatures*. Each Wickelfeature is a conjunctive, or context-sensitive, feature, capturing a feature of the central phoneme, a feature of the predecessor, and a feature of the successor.

Details of the Wickelfeature representation. For concreteness, we will now describe the details of the feature coding scheme we used. It contains several arbitrary properties, but it also captures the basic principles of coarse, conjunctive coding described in Chapter 3. First, we will describe the simple feature representation scheme we used for coding a single phoneme as a pattern of features without regard to its predecessor and successor. Then we describe how this scheme can be extended to code whole Wickelphones. Finally, we show how we "blur" this representation, to promote generalization further.

To characterize each phoneme, we devised the highly simplified feature set illustrated in Table 5. The purpose of the scheme was (a) to give as many of the phonemes as possible a distinctive code, (b) to allow code similarity to reflect the similarity structure of the phonemes

TABLE 5

CATEGORIZATION OF PHONEMES ON FOUR SIMPLE DIMENSIONS

		Place					
		Front		Middle		Back	
		V/L	U/S	V/L	U/S	V/L	U/S
Interrupted	*Stop*	b	p	d	t	g	k
	Nasal	m	-	n	-	N	-
Cont. Consonant	*Fric.*	v/D	f/T	z	s	Z/j	S/C
	Liq/SV	w/l	-	r	-	y	h
Vowel	*High*	E	i	O	^	U	u
	Low	A	e	I	a/α	W	*/o

Key: N = ng in *sing*; D = th in *the*; T = th in *with*; Z = z in *azure*; S = sh in *ship*; C = ch in *chip*; E = ee in *beet*; i = i in *bit*; O = oa in *boat*; ^ = u in *but* or schwa; U = oo in *boot*; u = oo in *book*; A = ai in *bait*; e = e in *bet*; I = i_e in *bite*; a = a in *bat*; α = a in *father*; W = ow in *cow*; * = aw in *saw*; o = o in *hot*.

in a way that seemed sufficient for our present purposes, and (c) to keep the number of different features as small as possible.

The coding scheme can be thought of as categorizing each phoneme on each of four dimensions. The first dimension divides the phonemes into three major types: interrupted consonants (stops and nasals), continuous consonants (fricatives, liquids, and semivowels), and vowels. The second dimension further subdivides these major classes. The interrupted consonants are divided into plain stops and nasals; the continuous consonants into fricatives and sonorants (liquids and semivowels are lumped together); and the vowels into high and low. The third dimension classifies the phonemes into three rough places of articulation—front, middle, and back. The fourth subcategorizes the consonants into voiced vs. voiceless categories and subcategorizes the vowels into long and short. As it stands, the coding scheme gives identical codes to six pairs of phonemes, as indicated by the duplicate entries in the cells of the table. A more adequate scheme could easily be constructed by increasing the number of dimensions and/or values on the dimensions.

Using the above code, each phoneme can be characterized by one value on each dimension. If we assigned a unit for each value on each dimension, we would need 10 units to represent the features of a single phoneme since two dimensions have three values and two have two

values. We could then indicate the pattern of these features that corresponds to a particular phoneme as a pattern of activation over the 10 units.

Now, one way to represent each Wickelphone would simply be to use three sets of feature patterns: one for the phoneme itself, one for its predecessor, and one for its successor. To capture the word-boundary marker, we would need to introduce a special eleventh feature. Thus, the Wickelphone $_{\#}\mathbf{k}_{\mathbf{a}}$ can be represented by

[(000) (00) (000) (00) 1]
[(100) (10) (001) (01) 0]
[(001) (01) (010) (01) 0].

Using this scheme, a Wickelphone could be represented as a pattern of activation over a set of 33 units.

However, there is one drawback with this. The representation is not sufficient to capture more than one Wickelphone at a time. If we add another Wickelphone, the representation gives us no way of knowing which features belong together.

We need a representation, then, that provides us with a way of determining which features go together. This is just the job that can be done with detectors for Wickelfeatures—triples of features, one from the central phoneme, one from the predecessor phoneme, and one from the successor phoneme.

Using this scheme, each detector would be activated when the word contained a Wickelphone containing its particular combination of three features. Since each phoneme of a Wickelphone can be characterized by 11 features (including the word-boundary feature) and each Wickelphone contains three phonemes, there are $11 \times 11 \times 11$ possible Wickelfeature detectors. Actually, we are not interested in representing phonemes that cross word boundaries, so we only need 10 features for the center phoneme.

Though this leaves us with a fairly reasonable number of units ($11 \times 10 \times 11$ or 1,210), it is still large by the standards of what will easily fit in available computers. However, it is possible to cut the number down still further without much loss of representational capacity since a representation using all 1,210 units would be highly redundant; it would represent each feature of each of the three phonemes 16 different times, one for each of the conjunctions of that feature with one of the four features of one of the other phonemes and one of the four features of the other.

To cut down on this redundancy and on the number of units required, we simply eliminated all those Wickelfeatures specifying values on two different dimensions of the predecessor and the

successor phonemes. We kept all the Wickelfeature detectors for all combinations of different values on the same dimension for the predecessor and successor phonemes. It turns out that there are 260 of these (ignoring the word-boundary feature), and each feature of each member of each phoneme triple is still represented four different times. In addition, we kept the 100 possible Wickelfeatures combining a preceding word-boundary feature with any feature of the main phoneme and any feature of the successor; and the 100 Wickelfeatures combining a following word boundary feature with any feature of the main phoneme and any feature of the successor. All in all then, we used only 460 of the 1,210 possible Wickelfeatures.

Using this representation, a verb is represented by a pattern of activation over a set of 460 Wickelfeature units. Each Wickelphone activates 16 Wickelfeature units. Table 6 shows the 16 Wickelfeature units activated by the Wickelphone ${}_{k}A_{m}$, the central Wickelphone in the word *came*. The first Wickelfeature is turned on whenever we have a Wickelphone in which the preceding contextual phoneme is an interrupted consonant, the central phoneme is a vowel, and the following phoneme is an interrupted consonant. This Wickelfeature is turned on for the Wickelphone ${}_{k}A_{m}$ since /k/ and /m/, the context phonemes, are both interrupted consonants and /A/, the central phoneme, is a vowel. This same Wickelfeature would be turned on in the

TABLE 6

THE SIXTEEN WICKELFEATURES FOR THE WICKELPHONE ${}_{k}A_{m}$

Feature	Preceding Context	Central Phoneme	Following Context
1	Interrupted	Vowel	Interrupted
2	Back	Vowel	Front
3	Stop	Vowel	Nasal
4	Unvoiced	Vowel	Voiced
5	Interrupted	Front	Vowel
6	Back	Front	Front
7	Stop	Front	Nasal
8	Unvoiced	Front	Voiced
9	Interrupted	Low	Interrupted
10	Back	Low	Front
11	Stop	Low	Nasal
12	Unvoiced	Low	Voiced
13	Interrupted	Long	Vowel
14	Back	Long	Front
15	Stop	Long	Nasal
16	Unvoiced	Long	Voiced

representation of ${}_{b}\mathbf{i}_{d}$, ${}_{p}\hat{}_{t}$, ${}_{m}\mathbf{a}_{p}$, and many other Wickelfeatures. Similarly, the sixth Wickelfeature listed in the table will be turned on whenever the preceding phoneme is made in the back, and the central and following phonemes are both made in the front. Again, this is turned on because /k/ is made in the back and /A/ and /m/ are both made in the front. In addition to ${}_{k}\mathbf{A}_{m}$ this feature would be turned on for the Wickelphones ${}_{g}\mathbf{i}_{v}$, ${}_{g}\mathbf{A}_{p}$, ${}_{k}\mathbf{A}_{p}$, and others. Similarly, each of the sixteen Wickelfeatures stands for a conjunction of three phonetic features and occurs in the representation of a large number of Wickelphones.

Now, words are simply lists of Wickelphones. Thus, words can be represented by simply turning on all of the Wickelfeatures in any Wickelphone of a word. Thus, a word with three Wickelphones (such as *came*, which has the Wickelphones ${}_{\#}\mathbf{k}_{A}$, ${}_{k}\mathbf{A}_{m}$, and ${}_{a}\mathbf{m}_{\#}$) will have at most 48 Wickelfeatures turned on. Since the various Wickelphones may have some Wickelfeatures in common, typically there will be less than 16 times the number of Wickelfeatures turned on for most words. It is important to note the temporal order is entirely implicit in this representation. All words, no matter how many phonemes in the word, will be represented by a subset of the 460 Wickelfeatures.

Blurring the Wickelfeature representation. The representational scheme just outlined constitutes what we call the *primary* representation of a Wickelphone. In order to promote faster generalization, we further blurred the representation. This is accomplished by turning on, in addition to the 16 primary Wickelfeatures, a randomly selected subset of the similar Wickelfeatures, specifically, those having the same value for the central feature and one of the two context phonemes. That is, whenever the Wickelfeature for the conjunction of phonemic features f_1, f_2, and f_3 is turned on, each Wickelfeature of the form $<?f_2f_3>$ and $<f_1f_2?>$ may be turned on as well. Here "?" stands for "any feature." This causes each word to activate a larger set of Wickelfeatures, allowing what is learned about one sequence of phonemes to generalize more readily to other similar but not identical sequences.

To avoid having too much randomness in the representation of a particular Wickelphone, we turned on the same subset of additional Wickelfeatures each time a particular Wickelphone was to be represented. Based on subsequent experience with related models (see Chapter 19), we do not believe this makes very much difference.

There is a kind of trade-off between the discriminability among the base forms of verbs that the representation provides and the amount of generalization. We need a representation which allows for rapid generalization while at the same time maintains adequate discriminability. We can manipulate this factor by manipulating the probability p that

any one of these similar Wickelfeatures will be turned on. In our simulations we found that turning on the additional features with fairly high probability (.9) led to adequate discriminability while also producing relatively rapid generalization.

Although the model is not completely immune to the possibility that two different words will be represented by the same pattern, we have encountered no difficulty decoding any of the verbs we have studied. However, we do not claim that Wickelfeatures necessarily capture all the information needed to support the generalizations we might need to make for this or other morphological processes. Some morphological processes might require the use of units that were further differentiated according to vowel stress or other potential distinguishing characteristics. All we claim for the present coding scheme is its sufficiency for the task of representing the past tenses of the 500 most frequent verbs in English and the importance of the basic principles of distributed, coarse (what we are calling blurred), conjunctive coding that it embodies (see Chapter 3).

Summary of the Structure of the Model

In summary, our model contained two sets of 460 Wickelfeature units, one set (the input units) to represent the base form of each verb and one set (the output units) to represent the past-tense form of each verb.

The model is tested by typing in an input phoneme string, which is translated by the fixed encoding network into a pattern of activation over the set of input units. Each active input unit contributes to the net input of each output unit, by an amount and direction (positive or negative) determined by the weight on the connection between the input unit and the output unit. The output units are then turned on or off probabilistically, with the probability increasing with the difference between the net input and the threshold, according to the logistic activation function. The output pattern generated in this way can be compared with various alternative possible output patterns, such as the correct past-tense form or some other possible response of interest, or can be used to drive the decoder network described in the Appendix.

The model is trained by providing it with pairs of patterns, consisting of the base pattern and the target, or correct, output. Thus, in accordance with common assumptions about the nature of the learning situation that faces the young child, the model receives only correct input from the outside world. However, it compares what it generates internally to the target output, and when it gets the wrong answer for a

particular output unit, it adjusts the strength of the connection between the input and the output units so as to reduce the probability that it will make the same mistake the next time the same input pattern is presented. The adjustment of connections is an extremely simple and *local* procedure, but it appears to be sufficient to capture what we know about the acquisition of the past tense, as we shall see in the next section.

THE SIMULATIONS

The simulations described in this section are concerned with demonstrating three main points:

- That the model captures the basic three-stage pattern of acquisition.
- That the model captures most aspects of differences in performance on different types of regular and irregular verbs.
- That the model is capable of responding appropriately to verbs it has never seen before, as well as to regular and irregular verbs actually experienced during training.

In the sections that follow we will consider these three aspects of the model's performance in turn.

The corpus of verbs used in the simulations consisted of a set of 506 verbs. All verbs were chosen from the Kucera and Francis (1967) word list and were ordered according to frequency of their gerund form. We divided the verbs into three classes: 10 high-frequency verbs, 410 medium-frequency verbs, and 86 low-frequency verbs. The ten highest frequency verbs were: *come* (/k^m/), *get* (/get/), *give* (/giv/), *look* (/luk/), *take* (/tAk/), *go* (/gO/), *have* (/hav/), *live* (/liv/), and *feel* (/fEl/). There is a total of 8 irregular and 2 regular verbs among the top 10. Of the medium-frequency verbs, 334 were regular and 76 were irregular. Of the low-frequency verbs, 72 were regular and 14 were irregular.

The Three-Stage Learning Curve

The results described in this and the following sections were obtained from a single (long) simulation run. The run was intended to capture

approximately the experience with past tenses of a young child picking up English from everyday conversation. Our conception of the nature of this experience is simply that the child learns first about the present and past tenses of the highest frequency verbs; later on, learning occurs for a much larger ensemble of verbs, including a much larger proportion of regular forms. Although the child would be hearing present and past tenses of all kinds of verbs throughout development, we assume that he is only able to learn past tenses for verbs that he has already mastered fairly well in the present tense.

To simulate the earliest phase of past-tense learning, the model was first trained on the 10 high-frequency verbs, receiving 10 cycles of training presentations through the set of 10 verbs. This was enough to produce quite good performance on these verbs. We take the performance of the model at this point to correspond to the performance of a child in Phase 1 of acquisition. To simulate later phases of learning, the 410 medium-frequency verbs were added to the first 10 verbs, and the system was given 190 more learning trials, with each trial consisting of one presentation of each of the 420 verbs. The responses of the model early on in this phase of training correspond to Phase 2 of the acquisition process; its ultimate performance at the end of 190 exposures to each of the 420 verbs corresponds to Phase 3. At this point, the model exhibits almost errorless performance on the basic 420 verbs. Finally, the set of 86 lower-frequency verbs were presented to the system and the transfer responses to these were recorded. During this phase, connection strengths were not adjusted. Performance of the model on these transfer verbs is considered in a later section.

We do not claim, of course, that this training experience exactly captures the learning experience of the young child. It should be perfectly clear that this training experience exaggerates the difference between early phases of learning and later phases, as well as the abruptness of the transition to a larger corpus of verbs. However, it is generally observed that the early, rather limited vocabulary of young children undergoes an explosive growth at some point in development (Brown, 1973). Thus, the actual transition in a child's vocabulary of verbs would appear quite abrupt on a time-scale of years so that our assumptions about abruptness of onset may not be too far off the mark.

Figure 4 shows the basic results for the high frequency verbs. What we see is that during the first 10 trials there is no difference between regular and irregular verbs. However, beginning on Trial 11 when the 410 midfrequency verbs were introduced, the regular verbs show better performance. It is important to notice that there is no interfering effect on the regular verbs as the midfrequency verbs are being learned. There is, however, substantial interference on the irregular verbs. This interference leads to a dip in performance on the irregular verbs.

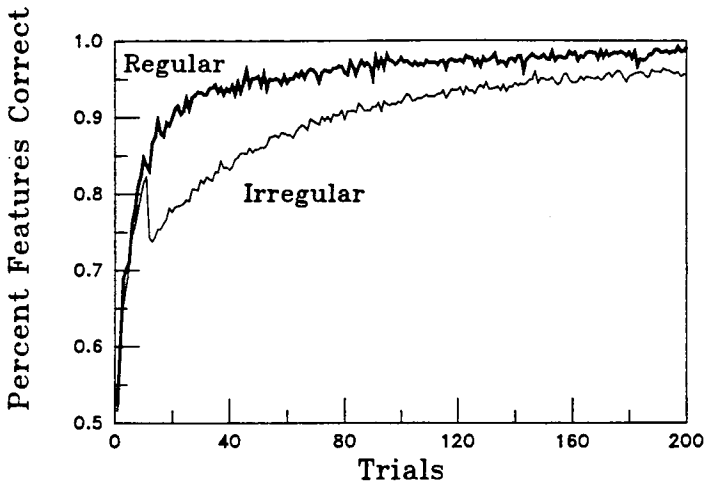


FIGURE 4. The percentage of correct features for regular and irregular high-frequency verbs as a function of trials.

Equality of performance between regular and irregular verbs is never again attained during the training period. This is the so-called U-shaped learning curve for the learning of the irregular past tense. Performance is high when only a few high-frequency, largely irregular verbs are learned, but then drops as the bulk of lower-frequency regular verbs are being learned.

We have thus far only shown that performance on high-frequency irregular verbs drops; we have not said anything about the nature of the errors. To examine this question, the response strength of various possible response alternatives must be compared. To do this, we compared the strength of response for several different response alternatives. We compared strengths for the correct past tense, the present, the base+ed and the past+ed. Thus, for example with the verb *give* we compared the response strength of /gAv/, /giv/, /givd/, and /gAvd/. We determined the response strengths by assuming that these response alternatives were competing to account for the features that were actually turned on in the output. The details of the competition mechanism, called a *binding network*, are described in the Appendix. For present purposes, suffice it to say that each alternative gets a score that represents the percentage of the total features that it accounts for. If two alternatives both account for a given feature, they divide the score for that feature in proportion to the number of features each accounts for uniquely. We take these response strengths to correspond roughly

to relative response probabilities, though we imagine that the actual generation of overt responses is accomplished by a different version of the binding network, described below. In any case, the total strength of all the alternatives cannot be greater than 1, and if a number of features are accounted for by none of the alternatives, the total will be less than 1.

Figure 5 compares the response strengths for the correct alternative to the combined strength of the regularized alternatives.[4] Note in the figure that during the first 10 trials the response strength of the correct alternative grows rapidly to over .5 while that of the regularized alternative drops from about .2 to .1. After the midfrequency verbs are introduced, the response strength for the correct alternative drops rapidly while the strengths of regularized alternatives jump up. From about Trials 11 through 30, the regularized alternatives together are stronger than the correct response. After about Trial 30, the strength of the correct response again exceeds the regularized alternatives and continues to grow throughout the 200-trial learning phase. By the end, the correct response is much the strongest with all other alternatives below .1.

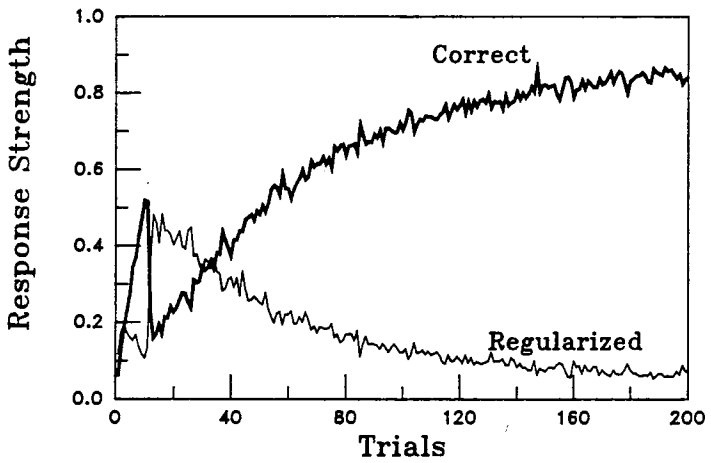


FIGURE 5. Response strengths for the high-frequency irregular verbs. The response strengths for the correct responses are compared with those for the regularized alternatives as a function of trials.

[4] Unless otherwise indicated, the regularized alternatives are considered the base+ed and past+ed alternatives. In a later section of the paper we shall discuss the pattern of differences between these alternatives. In most cases the base+ed alternative is much stronger than the past+ed alternative.

The rapidity of the growth of the regularized alternatives is due to the sudden influx of the medium-frequency verbs. In real life we would expect the medium-frequency verbs to come in somewhat more slowly so that the period of maximal regularization would have a somewhat slower onset.

Figure 6 shows the same data in a slightly different way. In this case, we have plotted the ratio of the correct response to the sum of the correct and regularized response strengths. Points on the curve below the .5 line are in the region where the regularized response is greater that the correct response. Here we see clearly the three stages. In the first stage, the first 10 trials of learning, performance on these high-frequency verbs is quite good. Virtually no regularization takes place. During the next 20 trials, the system regularizes and systematically makes errors on the verbs that it previously responded to correctly. Finally, during the remaining trials the model slowly eliminates the regularization responses as it approaches adult performance.

In summary, then, the model captures the three phases of learning quite well, as well as the gradual transition from Phase 2 to Phase 3. It does so without any explicit learning of rules. The regularization is the product of the gradual tuning of connection strengths in response

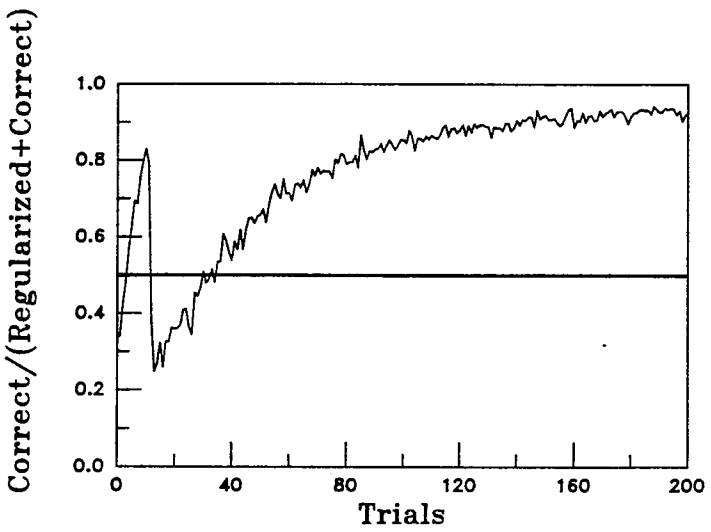


FIGURE 6. The ratio of the correct response to the sum of the correct and regularized response. Points on the curve below the .5 line are in the region where the regularized response is greater than the correct response.

to the predominantly regular correspondence exhibited by the medium-frequency words. It is not quite right to say that individual pairs are being stored in the network in any simple sense. The connection strengths the model builds up to handle the irregular forms do not represent these items in any separable way; they represent them in the way they must be represented to be stored along with the other verbs in the same set of connections.

Before discussing the implications of these kinds of results further, it is useful to look more closely at the kinds of errors made and at the learning rates of the medium-frequency regular and irregular verbs.

Learning the medium-frequency verbs. Figure 7A compares the learning curves for the regular verbs of high and medium frequency, and Figure 7B compares the learning curves for the corresponding groups of irregular verbs. Within only two or three trials the medium-frequency verbs catch up with their high-frequency counterparts. Indeed, in the case of the irregular verbs, the medium-frequency verbs seem to surpass the high-frequency ones. As we shall see in the following section, this results from the fact that the high-frequency verbs include some of the most difficult pairs to learn, including, for example, the *go/went* pair which is the most difficult to learn (aside from the verb *be*, this is the only verb in English in which the past and root form are completely unrelated). It should also be noted that even at this early stage of learning there is substantial generalization. Already, on Trial 11, the very first exposure to the medium-frequency verbs, between 65 and 75 percent of the features are produced correctly. Chance responding is only 50 percent. Moreover, on their first presentation, 10 percent more of the features of regular verbs are correctly responded to than irregular ones. Eventually, after 200 trials of learning, nearly all of the features are being correctly generated and the system is near asymptotic performance on this verb set. As we shall see below, during most of the learning period the difference between high- and medium-frequency verbs is not important. Rather, the differences between different classes of verbs is the primary determiner of performance. We now turn to a discussion of these different types.

Types of Regular and Irregular Verbs

To this point, we have treated regular and irregular verbs as two homogeneous classes. In fact, there are a number of distinguishable

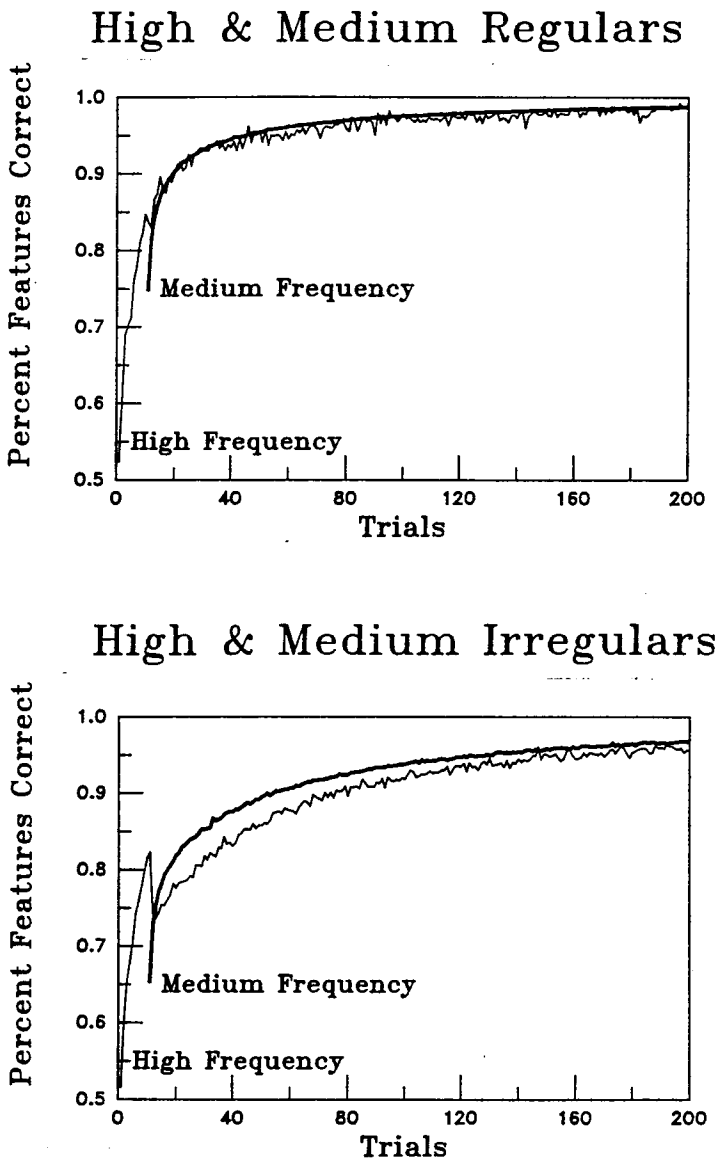


FIGURE 7. The learning curves for the high- and medium-frequency verbs.

types of regular and irregular verbs. Bybee and Slobin (1982) have studied the different acquisition patterns of the each type of verb. In this section we compare their results to the responses produced by our simulation model.

Bybee and Slobin divided the irregular verbs into nine classes, defined as follows: [5]

I. Verbs that do not change at all to form the past tense, e.g., *beat*, *cut*, *hit*.

II. Verbs that change a final /d/ to /t/ to form the past tense, e.g., *send/sent*, *build/built*.

III. Verbs that undergo an internal vowel change and also add a final /t/ or /d/, e.g., *feel/felt*, *lose/lost*, *say/said*, *tell/told*.

IV. Verbs that undergo an internal vowel change, delete a final consonant, and add a final /t/ or /d/, e.g., *bring/brought*, *catch/caught*.[6]

V. Verbs that undergo an internal vowel change whose stems end in a dental, e.g., *bite/bit*, *find/found*, *ride/rode*.

VIa. Verbs that undergo a vowel change of /i/ to /a/ e.g., *sing/sang*, *drink/drank*.

VIb. Verbs that undergo an internal vowel change of /i/ or /a/ to /ʌ/ e.g., *sting/stung*, *hang/hung*.[7]

VII. All other verbs that undergo an internal vowel change, e.g., *give/gave*, *break/broke*.

VIII. All verbs that undergo a vowel change and that end in a diphthongal sequence, e.g., *blow/blew*, *fly/flew*.

A complete listing by type of all of the irregular verbs used in our study is given in Table 7.

In addition to these types of irregular verbs, we distinguished three categories of regular verbs: (a) those ending in a vowel or voiced consonant, which take a /d/ to form the past tense; (b) those ending in a voiceless consonant, which take a /t/; and (c) those ending in /t/ or

---

[5] Criteria from Bybee and Slobin (1982, pp. 268-269).

[6] Following Bybee and Slobin, we included *buy/bought* in this class even though no final consonant is deleted.

[7] For many purposes we combine Classes VIa and VIb in our analyses.

TABLE 7

IRREGULAR VERBS

	Frequency		
Type	High	Medium	Low
I		beat fit set spread hit cut put	thrust bid
II		build send spend	bend lend
III	feel	deal do flee tell sell hear keep leave sleep lose mean say sweep	creep weep
IV	have make	think buy bring seek teach	catch
V	get	meet shoot write lead understand sit mislead bleed feed stand light find fight read meet hide hold ride	breed wind grind
VIa		drink ring sing swim	
VIb		drag hang swing	dig cling stick
VII	give take come	shake arise rise run become bear wear speak brake drive strike fall freeze choose	tear
VIII	go	throw blow grow draw fly know see	

/d/, which take a final /ˆd/ to form the past tense. The number of regular verbs in each category, for each of the three frequency levels, is given in Table 8.

Type I: No-change verbs. A small set of English verbs require no change between their present- and past-tense forms. One factor common to all such verbs is that they already end in /t/ or /d/. Thus, they superficially have the regular past-tense form—even in the present tense. Stemberger (1981) points out that it is common in inflectional languages not to add an additional inflection to base forms that already appear to have the inflection. Not all verbs ending in /t/ or /d/ show no change between present and past (in fact the majority of such verbs

TABLE 8

NUMBER OF REGULAR VERBS OF EACH TYPE

Type	Suffix	Example	Frequency		
			High	Medium	Low
End in dental	/ˆd/	start	0	94	13
End in voiceless consonant	/t/	look	1	64	30
End in voiced consonant or vowel	/d/	move	1	176	29

in English do show a change between present and past tense), but there is a reasonably large group—the Type I verbs of Bybee and Slobin—that do show this trend. Bybee and Slobin (1982) suggest that children learn relatively early on that past-tense verbs in English tend to end in /t/ or /d/ and thus are able to correctly respond to the no-change verbs rather early. Early in learning, they suggest, children also incorrectly generalize this "no-change rule" to verbs whose present and past tenses differ.

The pattern of performance just described shows up very clearly in data Bybee and Slobin (1982) report from an elicitation task with preschool children. In this task, preschoolers were given the present-tense form of each of several verbs and were asked to produce the corresponding past-tense form. They used the set of 33 verbs shown in Table 9.

The results were very interesting. Bybee and Slobin found that verbs not ending in *t/d* were predominately regularized and verbs ending in *t/d* were predominately used as no-change verbs. The number of occurrences of each kind is shown in Table 10. These preschool

TABLE 9

VERBS USED BY BYBEE & SLOBIN

Type of Verb	Verb List
Regular	walk smoke melt pat smile climb
Vowel change	drink break run swim throw meet shoot ride
Vowel change + *t/d*	do buy lose sell sleep help teach catch
No change	hit hurt set shut cut put beat
Other	go make build lend

TABLE 10

REGULAR AND NO CHANGE RESPONSES TO t/d AND OTHER VERBS
(Data from Bybee & Slobin, 1982)

Verb Ending	Regular Suffix	No Change
Not *t/d*	203	34
t/d	42	157

children have, at this stage, both learned to regularize verbs not ending in *t/d* and, largely, to leave verbs ending in *t/d* without an additional ending.

Interestingly, our simulations show the same pattern of results. The system learns both to regularize and has a propensity *not* to add an additional ending to verbs already ending in *t/d*. In order to compare the simulation results to the human data we looked at the performance of the same verbs used by Bybee and Slobin in our simulations. Of the 33 verbs, 27 were in the high- and medium-frequency lists and thus were included in the training set used in the simulation. The other six verbs (*smoke*, *catch*, *lend*, *pat*, *hurt* and *shut*) were either in the low-frequency sample or did not appear in our sample at all. Therefore, we will report on 27 out of the 33 verbs that Bybee and Slobin tested.

It is not clear what span of learning trials in our simulation corresponds best to the level of the preschoolers in Bybee and Slobin's experiment. Presumably the period during which regularization is occurring is best. The combined strength of the regularized alternatives exceeds correct response strength for irregulars from about Trial 11 through Trials 20 to 30 depending on which particular irregular verbs we look at. We therefore have tabulated our results over three different time ranges—Trials 11 through 15, Trials 16 through 20, and Trials 21 through 30. In each case we calculated the average strength of the regularized response alternatives and of the no-change response alternatives. Table 11 gives these strengths for each of the different time periods.

The simulation results show clearly the same patterns evident in the Bybee and Slobin data. Verbs ending in *t/d* always show a stronger no-change response and a weaker regularized response than those not ending in *t/d*. During the very early stages of learning, however, the regularized response is stronger than the no-change response—even if the verb does end with *t/d*. This suggests that the generalization that the past tense of *t/d* verbs is formed by adding /ˆd/ is stronger than the generalization that verbs ending in *t/d* should not have an ending

TABLE 11

AVERAGE SIMULATED STRENGTHS OF REGULARIZED AND NO-CHANGE RESPONSES

Time Period	Verb Ending	Regularized	No Change
11-15	not *t/d*	0.44	0.10
	t/d	0.35	0.27
16-20	not *t/d*	0.32	0.12
	t/d	0.25	0.35
21-30	not *t/d*	0.52	0.11
	t/d	0.32	0.41

added. However, as learning proceeds, this secondary generalization is made (though for only a subset of the *t/d* verbs, as we shall see), and the simulation shows the same interaction that Bybee and Slobin (1982) found in their preschoolers.

The data and the simulations results just described conflate two aspects of performance, namely, the tendency to make no-change *errors* with *t/d* verbs that are not no-change verbs and the tendency to make *correct* no-change responses to the *t/d* verbs that are no-change verbs. Though Bybee and Slobin did not report their data broken down by this factor, we can examine the results of the simulation to see whether in fact the model is making more no-change errors with *t/d* verbs for which this response is incorrect. To examine this issue, we return to the full corpus of verbs and consider the tendency to make no-change errors separately for irregular verbs other than Type I verbs and for regular verbs.

Erroneous no-change responses are clearly stronger for both regular and irregular *t/d* verbs. Figure 8A compares the strength of the erroneous no-change responses for irregular verbs ending in *t/d* (Types II and V) versus those not ending in *t/d* (Types III, IV, VI, VII, and VIII). The no-change response is erroneous in all of these cases. Note, however, that the erroneous no-change responses are stronger for the *t/d* verbs than for the other types of irregular verbs. Figure 8B shows the strength of erroneous no-change responses for regular verbs ending in *t/d* versus those not ending in *t/d*. Again, the response strength for the no-change response is clearly greater when the regular verb ends in a dental.

We also compared the regularization responses for irregular verbs whose stems end in *t/d* with irregulars not ending in *t/d*. The results are shown in Figure 8C. In this case, the regularization responses are initially stronger for verbs that do not end in *t/d* than for those that do.

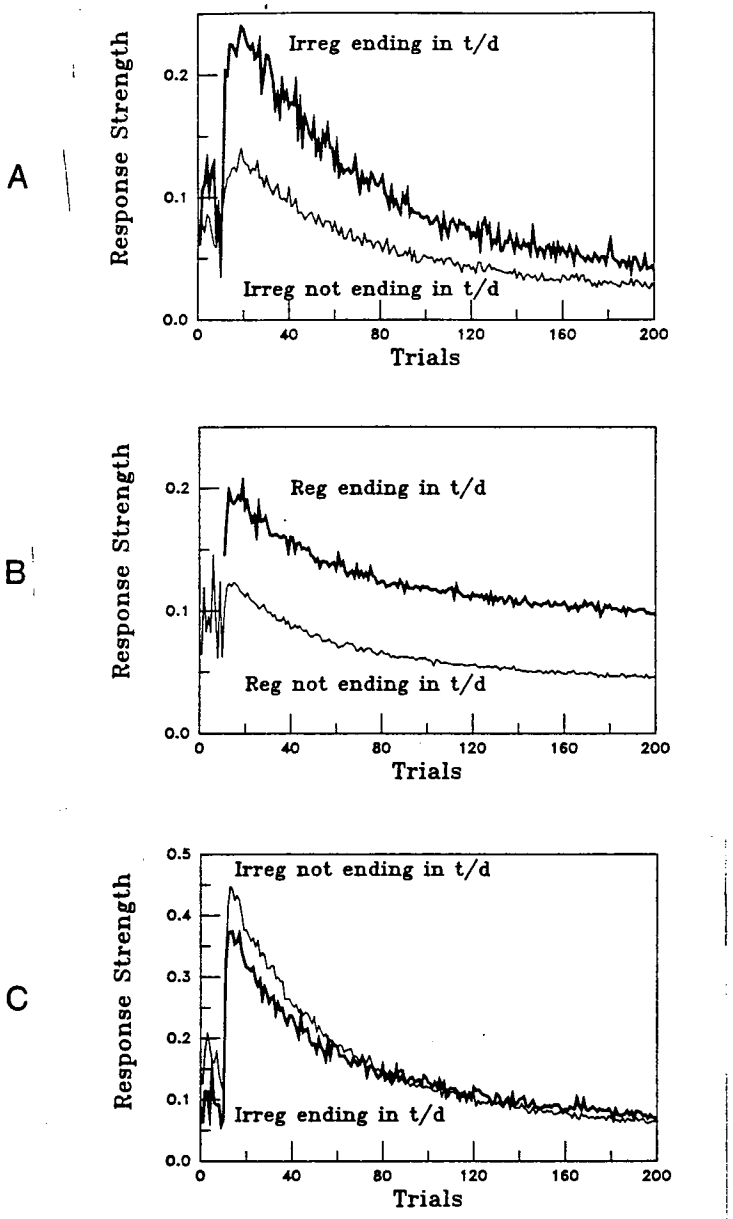


FIGURE 8. *A*: The strength of erroneous no-change responses for irregular verbs ending in a dental versus those not ending in a dental. *B*: The strength of erroneous no-change responses for regular verbs ending in a dental versus those not ending in a dental. *C*: The strength of erroneous regularization responses for irregular verbs ending in a dental versus those not ending in a dental.

Thus, we see that even when focusing only on erroneous responses, the system shows a greater propensity to respond with no change to *t/d* verbs, whether or not the verb is regular, and a somewhat greater tendency to regularize irregulars not ending in *t/d*.

There is some evidence in the literature on language acquisition that performance on Type I verbs is better sooner than for irregular verbs involving vowel changes—Types III through VIII. Kuczaj (1978) reports an experiment in which children were to judge the grammaticality of sentences involving past tenses. The children were given sentences involving words like *hit* or *hitted* or *ate* or *eated* and asked whether the sentences sounded "silly." The results, averaged over three age groups from 3;4 to 9;0 years, showed that 70 percent of the responses to the no-change verbs were correct whereas only 31 percent of the responses to vowel-change irregular verbs were correct. Most of the errors involved incorrect acceptance of a regularized form. Thus, the results show a clear difference between the verb types, with performance on the Type I verbs superior to that on Type III through VIII verbs.

The simulation model too shows better performance on Type I verbs than on any of the other types. These verbs show fewer errors than any of the other irregular verbs. Indeed the error rate on Type I verbs is equal to that on the most difficult of the regular verbs. Table 12 gives the average number of Wickelfeatures incorrectly generated (out of 460) at different periods during the learning processes for no-change (i.e., Type I) irregular verbs, vowel-change (i.e., Type III-VIII) irregular verbs, regular verbs ending in *t/d*, regular verbs not ending in *t/d*, and regular verbs ending in *t/d* whose stem is a CVC (consonant-vowel-consonant) monosyllable. The table clearly shows that throughout learning, fewer incorrect Wickelfeatures are generated for no-change verbs than for vowel-change verbs. Interestingly, the table

TABLE 12

AVERAGE NUMBER OF WICKELFEATURES INCORRECTLY GENERATED

Trial	Irregular Verbs		Regular Verbs		
Number	Type I	Types III-VIII	Ending in *t/d*	Not Ending in *t/d*	CV*t/d*
11-15	89.8	123.9	74.1	82.8	87.3
16-20	57.6	93.7	45.3	51.2	60.5
21-30	45.5	78.2	32.9	37.4	47.9
31-50	34.4	61.3	22.9	26.0	37.3
51-100	18.8	39.0	11.4	12.9	21.5
101-200	11.8	21.5	6.4	7.4	12.7

also shows that one subset of regulars are no easier than the Type I irregulars. These are the regular verbs which look on the surface most like Type I verbs, namely, the monosyllabic CVC regular verbs ending in *t/d*. These include such verbs as *bat*, *wait*, *shout*, *head*, etc. Although we know of no data indicating that people make more no-change errors on these verbs than on multisyllabic verbs ending in *t/d*, this is a clear prediction of our model. Essentially what is happening is that the model is learning that monosyllables ending in *t/d* sometimes take no additional inflection.[8] This leads to quicker learning of the no-change verbs relative to other irregular verbs and slower learning of regular verbs which otherwise look like no-change verbs. It should be noted that the two regular verbs employed by Bybee and Slobin which behaved like no-change verbs were both monosyllables. It would be interesting to see if whether no-change errors actually occur with verbs like *decide* or *devote*.

Types III-VIII: Vowel-change verbs. To look at error patterns on *vowel-change* verbs (Types III-VIII), Bybee and Slobin (1982) analyzed data from the spontaneous speech of preschoolers ranging from 1½ to 5 years of age. The data came from independent sets of data collected by Susan Ervin-Tripp and Wick Miller, by Dan Slobin, and by Zell Greenberg. In all, speech from 31 children involving the use of 69 irregular verbs was studied. Bybee and Slobin recorded the percentages of regularizations for each of the various types of vowel-change verbs. Table 13 gives the percentages of regularization by preschoolers, ranked from most to fewest erroneous regularizations. The results show that the two verb types which involve adding a *t/d* plus a vowel change (Types III and IV) show the least regularizations, whereas the verb type in which the present tense ends in a diphthong (Type VIII) shows by far the most regularization.

It is not entirely clear what statistic in our model best corresponds to the percentage of regularizations. It will be recalled that we collected response strength measures for four different response types for irregular verbs. These were the correct response, the no-change response, the base+ed regularization response, and the past+ed regularization response. If we imagine that no-change responses are, in general, difficult to observe in spontaneous speech, perhaps the measure that would be most closely related to the percentage of regularizations would be the ratio of the sum of the strengths of the regularization responses to

[8] Though the model does not explicitly encode number of syllables, monosyllabic words are distinguished from multisyllabic words by the fact that the former contain no Wickelphones of the form ${}_{V}C_{V}$. There are no no-change verbs in English containing such Wickelphones.

TABLE 13

PERCENTAGE OF REGULARIZATION BY PRESCHOOLERS
(Data from Bybee & Slobin, 1982)

Verb Type	Example	Percentage Regularizations
VIII	blew	80
VI	sang	55
V	bit	34
VII	broke	32
III	felt	13
IV	caught	10

the sum of the strengths of regularization responses and the correct response—that is,

$$\frac{(base{+}ed + past{+}ed)}{(base{+}ed + past{+}ed + correct)}.$$

As with our previous simulation, it is not entirely clear what portion of the learning curve corresponds to the developmental level of the children in this group. We therefore calculated this ratio for several different time periods around the period of maximal overgeneralization. Table 14 shows the results of these simulations.

The spread between different verb classes is not as great in the simulation as in the children's data, but the simulated rank orders show a

TABLE 14

STRENGTH OF REGULARIZATION RESPONSES RELATIVE TO CORRECT RESPONSES

	Data		Trials 11-15		Trials 16-20		Trials 21-30		Average Trials 11-30	
Rank Order	Type	Percent	Type	Ratio	Type	Ratio	Type	Ratio	Type	Ratio
1	VIII	80	VIII	.86	VIII	.76	VIII	.61	VIII	.71
2	VI	55	VII	.80	VII	.74	VII	.61	VII	.69
3	V	34	VI	.76	V	.60	IV	.48	V	.56
4	VII	32	V	.72	IV	.59	V	.46	IV	.56
5	III	13	IV	.69	III	.57	III	.44	III	.53
6	IV	10	III	.67	VI	.52	VI	.40	VI	.52

remarkable similarity to the results from the spontaneous speech of the preschoolers, especially in the earliest time period. Type VIII verbs show uniformly strong patterns of regularization whereas Type III and Type IV verbs, those whose past tense involves adding a *t/d* at the end, show relatively weak regularization responses. Type VI and Type VII verbs produce somewhat disparate results. For Type VI verbs, the simulation conforms fairly closely to the children's speech data in the earliest time period, but it shows rather less strength for regularizations of these verbs in the later time periods and in the average over Trials 11-30. For Type VII verbs, the model errs in the opposite direction: Here it tends to show rather greater strength for regularizations of these verbs than we see in the children's speech. One possible reason for these discrepancies may be the model's insensitivity to word frequency. Type VI verbs are, in fact, relatively low-frequency verbs, and thus, in the children's speech these verbs may actually be at a relatively earlier stage in acquisition than some of the more frequent irregular verbs. Type VII verbs are, in general, much more frequent—in fact, on the average they occur more than twice as often (in the gerund form) in the Kucera-Francis count than the Type VI verbs. In our simulations, all medium-frequency verbs were presented equally often and the distinction was not made. A higher-fidelity simulation including finer gradations of frequency variations among the verb types might lead to a closer correspondence with the empirical results. In any case, these verbs aside, the simulation seems to capture the major features of the data very nicely.

Bybee and Slobin attribute the pattern of results they found to factors that would not be relevant to our model. They proposed, for example, that Type III and IV verbs were more easily learned because the final *t/d* signaled to the child that they were in fact past tenses so the child would not have to rely on context as much in order to determine that these were past-tense forms. In our simulations, we found these verbs to be easy to learn, but it must have been for a different reason since the learning system was always informed as to what the correct past tense really was. Similarly, Bybee and Slobin argued that Type VIII verbs were the most difficult because the past and present tenses were so phonologically different that the child could not easily determine that the past and present tenses of these verbs actually go together. Again, our simulation showed Type VIII verbs to be the most difficult, but this had nothing to do with putting the past and present tense together since the model was always given the present and past tenses together.

Our model, then, must offer a different interpretation of Bybee and Slobin's findings. The main factor appears to be the degree to which the relation between the present and past tense of the verb is

idiosyncratic. Type VIII verbs are most difficult because the relationship between base form and past tense is most idiosyncratic for these verbs. Thus, the natural generalizations implicit in the population of verbs must be overcome for these verbs, and they must be overcome in a different way for each of them. A very basic aspect of the mapping from present to past tense is that most of the word, and in particular everything up to the final vowel, is unchanged. For regular verbs, all of the phonemes present in the base form are preserved in the past tense. Thus, verbs that make changes to the base form are going against the grain more than those that do not; the larger the changes, the harder they will be to learn. Another factor is that past tenses of verbs generally end in /t/ or /d/.

Verbs that violate the basic past-tense pattern are all at a disadvantage in the model, of course, but some suffer less than others because there are other verbs that deviate from the basic pattern in the same way. Thus, these verbs are less idiosyncratic than verbs such as *go/went*, *see/saw*, and *draw/drew* which represent completely idiosyncratic vowel changes. The difficulty with Type VIII verbs, then, is simply that, as a class, they are simply more idiosyncratic than other verbs. Type III and IV verbs (e.g., *feel/felt*, *catch/caught*), on the other hand, share with the vast bulk of the verbs in English the feature that the past tense involves the addition of a *t/d*. The addition of the *t/d* makes these verbs easier than, say, Type VII verbs (e.g., *come/came*) because in Type VII verbs the system must not only learn that there *is* a vowel change, but it must also learn that there *is not* an addition of *t/d* to the end of the verb.

Type VI verbs (*sing/sang*, *drag/drug*) are interesting from this point of view, because they involve fairly common subregularities not found in other classes of verbs such as those in Type V. In the model, the Type VI verbs may be learned relatively quickly because of this subregularity.

Types of regularization. We have mentioned that there are two distinct ways in which a child can regularize an irregular verb: The child can use the base+ed form or the past+ed form. Kuczaj (1977) has provided evidence that the proportion of past+ed forms increases, relative to the number of base+ed forms, as the child gets older. He found, for example, that the nine youngest children he studied had more base+ed regularizations than past+ed regularizations whereas four out of the five oldest children showed more past+ed than base+ed regularizations. In this section, we consider whether our model exhibits this same general pattern. Since the base form and the past-tense form are identical for Type I verbs, we restrict our analysis of this issue to Types II through VIII.

Figure 9 compares the average response strengths for base+ed and past+ed regularizations as a function of amount of training. The results of this analysis are more or less consistent with Kuczaj's findings. Early in learning, the base+ed response alternative is clearly the stronger of the two. As the system learns, however, the two come together so that by about 100 trials the base+ed and the past+ed response alternatives are roughly equally strong. Clearly, the simulations show that the percentage of regularizations that are past+ed increases with experience—just as Kuczaj found in children. In addition, the two curves come together rather late, consistent with the fact, reported by Kuczaj (1977), that these past+ed forms predominate for the most part in children who are exhibiting rather few regularization errors of either type. Of the four children exhibiting more past+ed regularizations, three were regularizing less than 12% of the time.

A closer look at the various types of irregular verbs shows that this curve is the average of two quite different patterns. Table 15 shows the overall percentage of regularization strength due to the base+ed alternative. It is clear from the table that the verbs fall into two general categories, those of Types III, IV, and VIII which have an overall preponderance of base+ed strength (the percentages are all above .5) and Types II, VII, V, and VI which show an overall preponderance of past+ed strength (the percentages are all well below .5). The major variable which seems to account for the ordering shown in the table is the amount the ending is changed in going from the base form to the

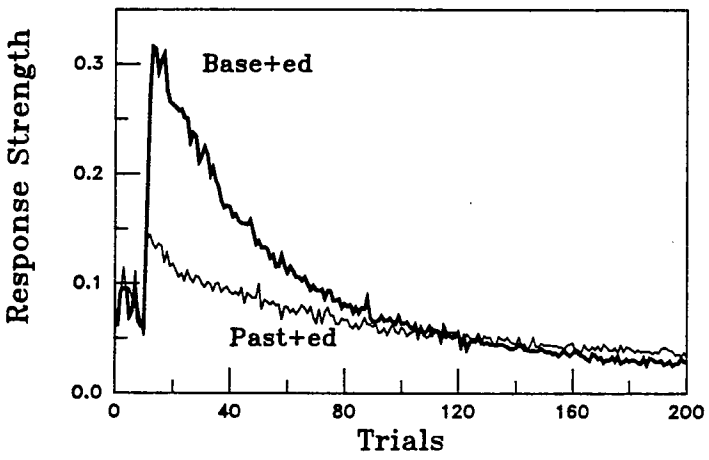


FIGURE 9. Average response strength for base+ed and past+ed responses for verb Types II through VIII.

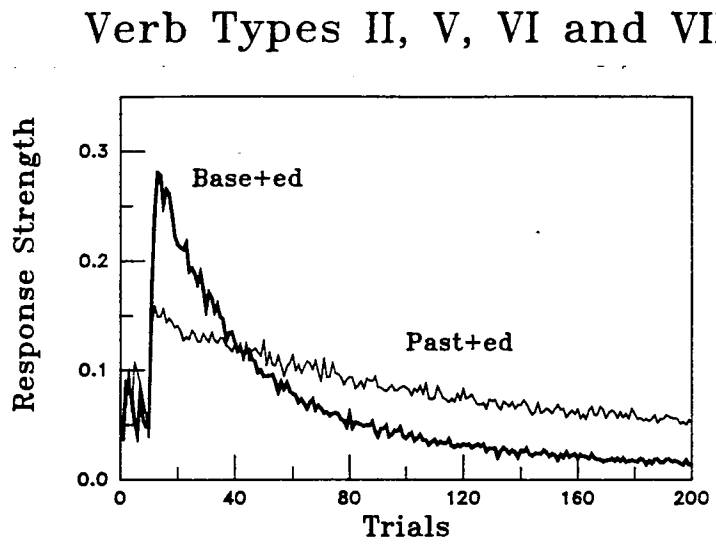
TABLE 15

PERCENTAGE OF REGULARIZATION STRENGTH DUE TO BASE+ED

Verb Type	Percent base+ed	Examples
III	0.77	sleep/slept
IV	0.69	catch/caught
VIII	0.68	see/saw
II	0.38	spend/spent
VII	0.38	come/came
V	0.37	bite/bit
VI	0.26	sing/sang

past-tense form. If the ending is changed little, as in *sing/sang* or *come/came*, the past+ed response is relatively stronger. If the past tense involves a greater change of the ending, such as *see/saw,* or *sleep/slept*, then the past+ed form is much weaker. Roughly, the idea is this: To form the past+ed for these verbs *two operations* must occur. The normal past tense must be created, and the regular ending must be appended. When these two operations involve very different parts of the verb, they can occur somewhat independently and both can readily occur. When, on the other hand, both changes occur to the same portion of the verb, they conflict with one another and a clear past+ed response is difficult to generate. The Type II verbs, which do show an overall preponderance of past+ed regularization strength, might seem to violate this pattern since it involves some change to the end in its past-tense form. Note, however, that the change is only a one feature change from /d/ to /t/ and thus is closer to the pattern of the verbs involving no change to the final phonemes of the verb. Figure 10A shows the pattern of response strengths to base+ed and past+ed regularizations for verb Types II, VII, V, and VI which involve relatively little change of the final phonemes from base to past form. Figure 10B shows the pattern of response strengths to base+ed and past+ed for verb Types III, IV, and VIII. Figure 10A shows very clearly the pattern expected from Kuczaj's results. Early in learning, base+ed responses are by far the strongest. With experience the past+ed response becomes stronger and stronger relative to the base+ed regularizations until, at about Trial 40, it begins to exceed it. Figure 10B shows a different pattern. For these verbs the past+ed form is weak throughout learning and never comes close to the base+ed regularization response. Unfortunately, Kuczaj did not present data on the relative frequency of the two types of regularizations separately for different verb types.

A



B

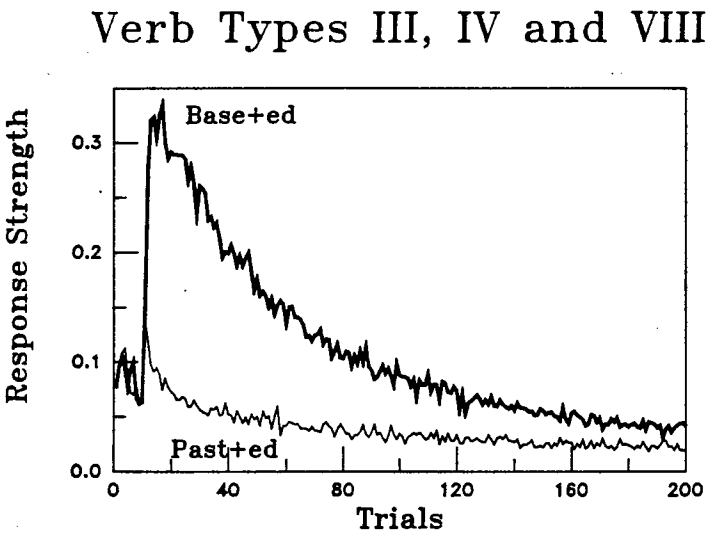


FIGURE 10. *A*: The pattern of response strengths to base+ed and past+ed regularizations for verb Types II, V, VI, and VII. *B*: The pattern of response strengths to base+ed and past+ed for verb Types III, IV, and VIII.

Thus for the present, this difference in type of regularization responses remains an untested prediction of the model.

Transfer to Novel Verbs

To this point we have only reported on the behavior of the system on verbs that it was actually taught. In this section, we consider the response of the model to the set of 86 low-frequency verbs which it never saw during training. This test allows us to examine how well the behavior of the model generalizes to novel verbs. In this section we also consider responses to different types of regular verbs, and we examine the model's performance in generating unconstrained responses.

Overall degree of transfer. Perhaps the first question to ask is how accurately the model generates the correct features of the new verbs. Table 16 shows the percentage of Wickelfeatures correctly generated, averaged over the regular and irregular verbs. Overall, the performance is quite good. Over 90 percent of the Wickelfeatures are correctly generated without any experience whatsoever with these verbs. Performance is, of course, poorer on the irregular verbs, in which the actual past tense is relatively idiosyncratic. But even there, almost 85 percent of the Wickelfeatures are correctly generated.

Unconstrained responses. Up until this point we have always proceeded by giving the model a set of response alternatives and letting it assign a response strength to each one. This allows us to get relative response strengths among the set of response alternatives we have provided. Of course, we chose as response alternatives those which we had reason to believe were among the strongest. There is the possibility, however, that the output of the model might actually favor some other, untested alternative some of the time. To see how well the output of the model is really doing at specifying correct past tenses or errors of the kind that children actually make, we must allow the model to choose among all possible strings of phonemes.

To do this, we implemented a second version of the binding network. This version is also described in the Appendix. Instead of a

TABLE 16

PROPORTION OF WICKELFEATURES CORRECTLY GENERATED

Regular	.92
Irregular	.84
Overall	.91

competition among alternative strings, it involves a competition among individual Wickelphone alternatives, coupled with mutual facilitation between mutually compatible Wickelphones such as $_{\#}\mathbf{k}_{\mathbf{A}}$ and $_{\mathbf{k}}\mathbf{A}_{\mathbf{m}}$.[9]

The results from the free-generation test are quite consistent wit h our expectations from the constrained alternative phase, though they did uncover a few interesting aspects of the model's performance that we had not anticipated. In our analysis of these results we have considered only responses with a strength of at least .2. Of the 86 test verbs, There were 65 cases in which exactly one of the alternatives exceeded .2. Of these, 55 were simple regularization responses, four were no-change responses, three involved double marking of regular verbs, (e.g., *type* was responded to with /tIpt^d/), and there was one case of a vowel change (e.g., *slip/slept*). There were 14 cases in which two alternatives exceeded threshold and one case in which three exceeded threshold. Finally, in six cases, no response alternative exceeded threshold. This occurred with the regular verbs *jump*, *pump*, *soak*, *warm*, *trail*, and *glare*. In this case there were a number of alternatives, including the correct past-tense form of each of these verbs, competing with a response strength of about .1.

Table 17 shows the responses generated for the 14 irregular verbs. The responses here are very clear. All of the above-threshold responses made to an irregular verb were either regularization responses, no-change responses (to Type I and V verbs as expected) or correct vowel-change generalizations. The fact that *bid* is correctly generated as the past for *bid,* that *wept* is correctly generated as the past for *weep*, and that *clung* is correctly generated as a past tense for *cling* illustrates that the system is not only sensitive to the major regular past-tense pattern, but is sensitive to the subregularities as well. It should also be noted that the no-change responses to the verbs *grind* and *wind* occurs on monosyllabic Type V verbs ending in *t/d,* again showing evidence of a role for this subregularity in English past-tense formation.

Of the 72 regular verbs in our low-frequency sample, the six verbs mentioned above did not have any response alternatives above threshold. On 48 of the remaining 66 regular verbs, the. only response exceeding threshold was the correct one. The threshold responses to the remaining 18 verbs are shown in Table 18.

---

[9] The major problem with this method of generating responses is that it is tremendously computer intensive. Had we used this method to generate responses throughout the learning phase, we estimate that it would have taken *over three years* of computer time to complete the learning phase alone! This compares to the 260 hours of computer time the learning phase took with the response alternatives supplied. It took about 28 hours to complete the response generation process in testing just the 86 low-frequency verbs used in this section of the study.

TABLE 17

THE MODEL'S RESPONSES TO UNFAMILIAR LOW-FREQUENCY IRREGULAR VERBS

Verb Type	Presented Word	Phonetic Input	Phonetic Response	English Rendition	Response Strength
I	bid	/bid/	/bid/	(bid)	0.55
	thrust	/ΤrˆSt/	/ΤrˆStˆd/	(thrusted)	0.57
II	bend	/bend/	/bendˆd/	(bended)	0.28
	lend	/lend/	/lendˆd/	(lended)	0.70
III	creep	/krEp/	/krEpt/	(creeped)	0.51
	weep	/wEp/	/wEpt/	(weeped)	0.34
			/wept/	(wept)	0.33
IV	catch	/kaC/	/kact/	(catched)	0.67
V	breed	/brEd/	/brEdˆd/	(breeded)	0.48
	grind	/grInd/	/grInd/	(grind)	0.44
	wind	/wInd/	/wInd/	(wind)	0.37
VI	cling	/klIN/	/klINd/	(clinged)	0.28
			/klˆN/	(clung)	0.23
	dig	/dig/	/digd/	(digged)	0.22
	stick	/stik/	/stikt/	(sticked)	0.53
VII	tear	/ter/	/terd/	(teared)	0.90

Note that for 12 of the 18 verbs listed in the table, the correct response is above threshold. That means that of the 66 regular verbs to which any response at all exceeded threshold, the correct response exceeded threshold in 60 cases. It is interesting to note, also, that the model never chooses the incorrect variant of the regular past tense. As shown in Table 8, verbs ending in a /t/ or /d/ take /ˆd/ in the past tense; verbs ending in unvoiced consonants take /t/, and verbs ending in vowels or voiced consonants take /d/. On no occasion does the model assign a strength greater than .2 an incorrect variant of the past tense. Thus, the model has clearly learned the substructure of the regular correspondence and adds the correct variant to all different types of base forms. These results clearly demonstrate that the model acts in accordance with the regular pattern for English verbs and that it can apply this pattern with a high level of success to novel as well as familiar verbs.

In addition to the regular responses, five of the responses were no-change responses. In three cases the no-change response was to a verb ending in *t/d*. Four of the responses followed the pattern of Type III verbs, modifying the vowel and adding a final /t/. Thus, for example,

TABLE 18

SYSTEM RESPONSES TO UNFAMILIAR LOW-FREQUENCY REGULAR VERBS

Verb Type	Presented Word	Phonetic Input	Phonetic Response	English Rendition	Response Rendition
End in *t/d*	guard	/gɑrd/	/gɑrd/	(guard)	0.29
			/gɑrd^d/	(guarded)	0.26
	kid	/kid/	/kid/	(kid)	0.39
			/kid^d/	(kidded)	0.24
	mate	/mAt/	/mAt^d/	(mated)	0.43
			/mAd^d/	(maded)	0.23
	squat	/skw*t/	/skw*t^d/	(squated)	0.27
			/skw*t/	(squat)	0.22
			/skw*kt/	(squawked)	0.21
End in unvoiced consonant	carp	/kɑrp/	/kɑrpt/	(carped)	0.28
			/kɑpt^d/	(carpted)	0.21
	drip	/drip/	/dript^d/	(dripted)	0.28
			/dript/	(dripped)	0.22
	map	/map/	/mapt^d/	(mapted)	0.24
			/mapt/	(mapped)	0.22
	shape	/sAp/	/sApt/	(shaped)	0.43
			/sipt/	(shipped)	0.27
	sip	/sip/	/sipt/	(sipped)	0.42
			/sept/	(sepped)	0.28
	slip	/slip/	/slept/	(slept)	0.40
	smoke	/smOk/	/smOkt^d/	(smokted)	0.29
			/smOk/	(smoke)	0.22
	snap	/snap/	/snapt^d/	(snapted)	0.40
	step	/step/	/stept^d/	(stepted)	0.59
	type	/tɪp/	/tɪpt^d/	(typted)	0.33
End in voiced consonant or vowel	brown	/brwn/	/brwnd/	(browned)	0.46
			/br*nd/	(brawned)	0.39
	hug	/h^g/	/h^g/	(hug)	0.59
	mail	/mA^l/	/mA^ld/	(mailed)	0.38
			/memb^ld/	(membled)	0.23
	tour	/tUr/	/tUrd^r/	(toureder)	0.31
			/tUrd/	(toured)	0.25

we have the past of *sip* rendered as *sept*, presumably on the model of *sleep/slept*, *keep/kept*, *sweep/swept*, etc. Interestingly, three of the four cases involved verbs whose base form ended in /p/ just as in the models listed above. Even though these last responses are, strictly

speaking, incorrect, they all indicate a sensitivity to the regular and subregular patterns of the English past tense.

Perhaps the most surprising result evident from the table is the occurrence of a double past marker on the responses to seven of the verbs. Although we know of no references to this phenomenon in the literature, we expect that children (and adults) do occasionally make this kind of error. It is interesting, and possibly significant, that all seven of these responses occurred to verbs whose correct past tense is the addition of a /t/. It would be interesting to see whether children's errors of this type follow a similar pattern.

Finally, there were just four responses that involved the addition or modification of consonants. These were *maded* as a past tense of *mate*, *squawked* as a past tense for *squat, membled* as a past tense for *mail*, and *toureder* as a past tense for *tour*. It is unlikely that humans would make these errors, especially the last two, but these responses are, for the most part, near threshold. Furthermore, it seems likely that many of these responses could be filtered out if the model incorporated an auto-associative network of connections among the output units. Such a network could be used to clean up the output pattern and would probably increase the tendency of the model to avoid bizarre responses. Unfortunately, we have not yet had the chance to implement this suggestion.

Summary. The system has clearly learned the essential characteristics of the past tense of English. Not only can it respond correctly to the 460 verbs that it was taught, but it is able to generalize and transfer rather well to the unfamiliar low-frequency verbs that had never been presented during training. The system has learned about the conditions in which each of the three regular past-tense endings are to be applied, and it has learned not only the dominant, regular form of the past tense, but many of the subregularities as well.

It is true that the model does not act as a perfect rule-applying machine with novel past-tense forms. However, it must be noted that people—or at least children, even in early grade-school years—are not perfect rule-applying machines either. For example, in Berko's classic (1958) study, though her kindergarten and first-grade subjects did often produce the correct past forms of novel verbs like *spow*, *mott*, and *rick*, they did not do so invariably. In fact, the rate of regular past-tense forms given to Berko's novel verbs was only 51 percent.[10] Thus, we see

[10] Unfortunately, Berko included only one regular verb to compare to her novel verbs. The verb was *melt*. Children were 73 percent correct on this verb. The two novel verbs that required the same treatment as *melt* (*mott* and *bodd*) each received only 33 percent correct responses.

little reason to believe that our model's "deficiencies" are significantly greater than those of native speakers of comparable experience.

CONCLUSIONS

We have shown that our simple learning model shows, to a remarkable degree, the characteristics of young children learning the morphology of the past tense in English. We have shown how our model generates the so-called U-shaped learning curve for irregular verbs and that it exhibits a tendency to overgeneralize that is quite similar to the pattern exhibited by young children. Both in children and in our model, the verb forms showing the most regularization are pairs such as *know/knew* and *see/saw*, whereas those showing the least regularization are pairs such as *feel/felt* and *catch/caught.* Early in learning, our model shows the pattern of more no-change responses to verbs ending in *t/d* whether or not they are regular verbs, just as young children do. The model, like children, can generate the appropriate regular past-tense form to unfamiliar verbs whose base form ends in various consonants or vowels. Thus, the model generates an /ˆd/ suffix for verbs ending in *t/d,* a /t/ suffix for verbs ending in an unvoiced consonant, and a /d/ suffix for verbs ending in a voiced consonant or vowel.

In the model, as in children, different past-tense forms for the same word can coexist at the same time. On rule accounts, such *transitional* behavior is puzzling and difficult explain. Our model, like human children, shows an relatively larger proportion of past+ed regularizations later in learning. Our model, like learners of English, will sometimes generate past-tense forms to novel verbs which show sensitivities to the subregularities of English as well as the major regularities. Thus, the past of *cring* can sometimes be rendered *crang* or *crung.* In short, our simple learning model accounts for all of the major features of the acquisition of the morphology of the English past tense.

In addition to our ability to account for the major *known* features of the acquisition process, there are also a number of predictions that the model makes which have yet to be reported. These include:

- We expect relatively more past+ed regularizations to irregulars whose correct past form *does not* involve a modification of the final phoneme of the base form.

- We expect that early in learning, a no-change response will occur more frequently to a CVC monosyllable ending in *t/d* than to a more complex base verb form.

- We expect that the double inflection responses (/dript^d/) will occasionally be made by native speakers and that they will occur more frequently to verbs whose stem is ends in /p/ or /k/.

The model is very rich and there are many other more specific predictions which can be derived from it and evaluated by a careful analysis of acquisition data.

We have, we believe, provided a distinct alternative to the view that children learn the rules of English past-tense formation in any explicit sense. We have shown that a reasonable account of the acquisition of past tense can be provided without recourse to the notion of a "rule" as anything more than a *description* of the language. We have shown that, for this case, there is no *induction problem.* The child need not figure out what the rules are, nor even that there are rules. The child need not decide whether a verb is regular or irregular. There is no question as to whether the inflected form should be stored directly in the lexicon or derived from more general principles. There isn't even a question (as far as generating the past-tense form is concerned) as to whether a verb form is one encountered many times or one that is being generated for the first time. A uniform procedure is applied for producing the past-tense form in every case. The base form is supplied as input to the past-tense network and the resulting pattern of activation is interpreted as a phonological representation of the past form of that verb. This is the procedure whether the verb is regular or irregular, familiar or novel.

In one sense, every form must be considered as being derived. In this sense, the network can be considered to be one large rule for generating past tenses from base forms. In another sense, it is possible to imagine that the system simply stores a set of rote associations between base and past-tense forms with novel responses generated by "on-line" generalizations from the stored exemplars.

Neither of these descriptions is quite right, we believe. Associations are simply stored in the network, but because we have a *superpositional* memory, similar patterns blend into one another and reinforce each other. If there were no similar patterns (i.e., if the featural representations of the base forms of verbs were orthogonal to one another) there would be no generalization. The system would be unable to generalize and there would be no regularization. It is statistical relationships among the base forms themselves that determine the pattern of responding. The network merely reflects the statistics of the featural representations of the verb forms.

We chose the study of acquisition of past tense in part because the phenomenon of regularization is an example often cited in support of

the view that children do respond according to general rules of language. Why otherwise, it is sometimes asked, should they generate forms that they have never heard? The answer we offer is that they do so because the past tenses of similar verbs they are learning show such a consistent pattern that the generalization from these similar verbs outweighs the relatively small amount of learning that has occurred on the irregular verb in question. We suspect that essentially similar ideas will prove useful in accounting for other aspects of language acquisition. We view this work on past-tense morphology as a step toward a revised understanding of language knowledge, language acquisition, and linguistic information processing in general.

ACKNOWLEDGMENTS

This research was supported by ONR Contracts N00014-82-C-0374, NR 667-483 and N00014-79-C-0323, NR 667-437, by a grant from the System Development Foundation, and by a Research Scientist Career Development Award MH00385 to the second author from the National Institute of Mental Health.

APPENDIX

One important aspect of the Wickelfeature representation is that it completely suppressed the temporal dimension. Temporal information is stored implicitly in the feature pattern. This gives us a representational format in which phonological forms of arbitrary length can be represented. It also avoids an a priori decision as to which part of the verb (beginning, end, center, etc.) contains the past-tense inflection. This grows out of the learning process. Unfortunately, it has its negative side as well. Since phonological forms *do* contain temporal information, we need to have a method of converting from the Wickelfeature representation into the time domain—in short, we need a decoding network which converts from the Wickelfeature representation to either the Wickelphone or a phonological representational format. Since we have probabilistic units, this decoding process must be able to work in the face of substantial noise. To do this we devised a special sort of decoding network which we call a *binding network*. Roughly speaking, a binding network is a scheme whereby a number of units *compete* for a set of available features—finally attaining a strength that is proportional to the number of features the units account for. We proceed by first describing the idea behind the binding network, then describing its application to produce the set of Wickelphones implicit in the Wickelfeature representation, and finally to produce the set of phonological strings implicit in the Wickelfeatures.

Binding Networks

The basic idea is simple. Imagine that there are a set of input features and a set of output features. Each output feature is consistent with certain of the input features, inconsistent with certain other of the input features, and neutral about still other of the input features. The idea is to find a set of output features that accounts for as many as possible of the output features while minimizing the number of input features accounted for by more than one output feature. Thus, we want each of the output features to *compete* for input features. The more input features it *captures*, the stronger its position in the competition and the more claim it has on the features it accounts for. Thus consider the case in which the input features are Wickelfeatures and the output features are Wickelphones. The Wickelphones compete among one another for the available Wickelfeatures. Every time a particular Wickelphone "captures" a particular Wickelfeature, that input feature no

longer provides support for other Wickelphones. In this way, the system comes up with a set of more or less nonoverlapping Wickelphones which account for as many as possible of the available Wickelfeatures. This means that if two Wickelphones have many Wickelfeatures in common (e.g., $\mathbf{\hat{k}_m}$ and $\mathbf{_kA_m}$) but one of them accounts for more features than the other, the one that accounts for the most features will remove nearly all of the support for the very similar output feature which accounts for few if any input features uniquely. The binding network described below has the property that if two output units are competing for a set of input features, each will attain a strength proportional to the number of input features uniquely accounted for by that output feature divided by the total number of input features uniquely accounted for by any output feature.

This is accomplished by a network in which each input unit has a fixed amount of activation (in our case we assumed that it had a total activation value of 1) to be distributed among the output units consistent with that input feature. It distributes its activation in proportion to the strength of the output feature to which it is connected. This is thus a network with a dynamic weight. The weight from input unit j to output unit i is thus given by

$$w_{ij} = \frac{a_i}{\sum_{k_j} a_{k_j}}$$

where k_j ranges over the set of output units consistent with input units j. The total strength of output unit k at time t is a linear function of its inputs at time $t-1$ and is thus given by

$$a_k(t) = \sum_{j_k} i_{j_k} w_{kj_k}(t) = \frac{\sum_{j_k} i_{j_k} a_k(t-1)}{\sum_{l_{j_k}} a_{l_{j_k}}(t-1)}$$

where j_k ranges over the set of input features consistent with output feature k, l_{j_k} ranges over the set of output features consistent with input feature j_k, and i_j takes on value 1 if input feature j is present and is 0 otherwise.

We used the binding network described above to find the set of Wickelphones which gave optimal coverage to the Wickelfeatures in the input. The procedure was quite effective. We used as the set of output all of the Wickelphones that occurred anywhere in any of the 500 or so verbs we studied. We found that the actual Wickelphones were always the strongest when we had 80 percent or more of the correct Wickelfeatures. Performance dropped off as the percentage of correct

Wickelfeatures dropped. Still when as few as 50 percent of the Wickelfeatures were correct, the correct Wickelphones were still the strongest most of the time. Sometimes, however, a Wickelphone not actually in the input would become strong and push out the "correct" Wickelphones. If we added the constraint that the Wickelphones must fit together to form an entire string (by having output features activate features that are consistent neighbors), we found that more than 60 percent of correct Wickelfeatures lead to the correct output string more than 90 percent of the time.

The binding network described above is designed for a situation in which there is a set of input features that is to be divided up among a set of output features. In this case, features that are present, but not required for a particular output feature play no role in the evaluation of the output feature. Suppose, however, that we have a set of alternative output features, one of which is supposed to account for the entire pattern. In this case, input features that are present, but not consistent, with a given output feature must count against that output feature. One solution to this is to have input units *excite* consistent output units according the the rule given above and to *inhibit* inconsistent output units. In the case in which we tried to construct the entire phonological string directly from a set of Wickelfeatures we used the following activation rule:

$$a_k(t) = \sum_{j_k} i_{j_k} w_{kj_k}(t) - \sum_{l_k} i_{l_k}$$

where l_k indexes the input features that are inconsistent with output feature k. In this case, we used as output features all of the strings of less than 20 phonemes which could be generated from the set of Wickelphones present in the entire corpus of verbs. This is the procedure employed to produce responses to the lowest frequency verbs as shown in Tables 17 and 18.