Cognition, 28 (1988) 73–193

On language and connectionism: Analysis of a parallel distributed processing model of language acquisition*

STEVEN PINKER
Massachusetts Institute of Technology

ALAN PRINCE
Brandeis University

Abstract

*Does knowledge of language consist of mentally-represented rules? Rumelhart and McClelland have described a connectionist (parallel distributed processing) model of the acquisition of the past tense in English which successfully maps many stems onto their past tense forms, both regular (*walk/walked*) and irregular (*go/went*), and which mimics some of the errors and sequences of development of children. Yet the model contains no explicit rules, only a set of neuron-style units which stand for trigrams of phonetic features of the stem, a set of units which stand for trigrams of phonetic features of the past form, and an array of connections between the two sets of units whose strengths are modified during learning. Rumelhart and McClelland conclude that linguistic rules may be merely convenient approximate fictions and that the real causal processes in language use and acquisition must be characterized as the transfer of activation levels among units and the modification of the weights of their connections. We analyze both the linguistic and the developmental assumptions of the model in detail and discover that (1) it cannot represent certain words, (2) it cannot learn many rules, (3) it can learn rules found in no human language, (4) it cannot explain morphological and phonological regularities, (5) it cannot ex-*

*The authors contributed equally to this paper and list their names in alphabetical order. We are grateful to Jane Grimshaw and Brian MacWhinney for providing transcripts of children's speech from the Brandeis Longitudinal Study and the Child Language Data Exchange System, respectively. We also thank Tom Bever, Jane Grimshaw, Stephen Kosslyn, Dan Slobin, an anonymous reviewer from *Cognition*, and the Boston Philosophy and Psychology Discussion Group for their comments on earlier drafts, and Richard Goldberg for his assistance. Preparation of this paper was supported by NSF grant IST-8420073 to Jane Grimshaw and Ray Jackendoff of Brandeis University, by NIH grant HD 18381-04 and NSF grant 85-18774 to Steven Pinker, and by a grant from the Alfred P. Sloan Foundation to the MIT Center for Cognitive Science. Requests for reprints may be sent to Steven Pinker at the Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, U.S.A. or Alan Prince at the Linguistics and Cognitive Science Program, Brown 125, Brandeis University, Waltham MA 02254, U.S.A.

plain the differences between irregular and regular forms, (6) it fails at its assigned task of mastering the past tense of English, (7) it gives an incorrect explanation for two developmental phenomena: stages of overregularization of irregular forms such as bringed, *and the appearance of doubly-marked forms such as* ated *and (8) it gives accounts of two others (infrequent overregularization of verbs ending in* t/d, *and the order of acquisition of different irregular subclasses) that are indistinguishable from those of rule-based theories. In addition, we show how many failures of the model can be attributed to its connectionist architecture. We conclude that connectionists' claims about the dispensability of rules in explanations in the psychology of language must be rejected, and that, on the contrary, the linguistic and developmental facts provide good evidence for such rules.*

> If design govern in a thing so small.
> *Robert Frost*

1. Introduction

The study of language is notoriously contentious, but until recently, researchers who could agree on little else have all agreed on one thing: that linguistic knowledge is couched in the form of rules and principles. This conception is consistent with—indeed, is one of the prime motivations for—the "central dogma" of modern cognitive science, namely that intelligence is the result of processing symbolic expressions. To understand language and cognition, according to this view, one must break them up into two aspects: the rules or symbol-manipulating processes capable of generating a domain of intelligent human performance, to be discovered by examining systematicities in people's perception and behavior, and the elementary symbol-manipulating mechanisms made available by the information-processing capabilities of neural tissue, out of which the rules or symbol-manipulating processes would be composed (see, e.g., Chomsky, 1965; Fodor, 1968, 1975; Marr, 1982; Minsky, 1963; Newell & Simon, 1961; Putnam, 1960; Pylyshyn, 1984).

One of the reasons this strategy is inviting is that we know of a complex intelligent system, the computer, that can only be understood using this algorithm-implementation or software-hardware distinction. And one of the reasons that the strategy has remained compelling is that it has given us precise, revealing, and predictive models of cognitive domains that have required few assumptions about the underlying neural hardware other than that it makes available some very general elementary processes of comparing and transforming symbolic expressions.

Of course, no one believes that cognitive models explicating the systematicities in a domain of intelligence can fly in the face of constraints provided by the operations made available by neural hardware. Some early cognitive models have assumed an underlying architecture inspired by the historical and technological accidents of current computer design, such as rapid reliable serial processing, limited bandwidth communication channels, or rigid distinctions between registers and memory. These assumptions are not only inaccurate as descriptions of the brain, composed as it is of slow, noisy and massively interconnected units acting in parallel, but they are unsuited to tasks such as vision where massive amounts of information must be processed in parallel. Furthermore, some cognitive tasks seem to require mechanisms for rapidly satisfying large sets of probabilistic constraints, and some aspects of human performance seem to reveal graded patterns of generalization to large sets of stored exemplars, neither of which is easy to model with standard serial symbol-matching architectures. And progress has sometimes been stymied by the difficulty of deciding among competing models of cognition when one lacks any constraints on which symbol-manipulating processes the neural hardware supplies "for free" and which must be composed of more primitive processes.

1.1. Connectionism and symbol processing

In response to these concerns, a family of models of cognitive processes originally developed in the 1950s and early 1960s has received increased attention. In these models, collectively referred to as "Parallel Distributed Processing" ("PDP") or "Connectionist" models, the hardware mechanisms are networks consisting of large numbers of densely interconnected units, which correspond to concepts (Feldman & Ballard, 1982) or to features (Hinton, McClelland, & Rumelhart, 1981). These units have activation levels and they transmit signals (graded or 1–0) to one another along weighted connections. Units "compute" their output signals through a process of weighting each of their input signals by the strength of the connection along which the signal is coming in, summing the weighted input signals, and feeding the result into a nonlinear output function, usually a threshold. Learning consists of adjusting the strengths of connections and the threshold-values, usually in a direction that reduces the discrepancy between an actual output in response to some input and a "desired" output provided by an independent set of "teaching" inputs. In some respects, these models are thought to resemble neural networks in meaningful ways; in others, most notably the teaching and learning mechanisms, there is no known neurophysiological analogue, and some authors are completely agnostic about how the units and connections are neurally instantiated. ("Brain-style modeling" is the noncommittal term used by Rumelhart & McClelland, 1986a.) The computations underlying cognitive processes occur when a set of input units in a network is turned on in a pattern that corresponds in a fixed way to a stimulus or internal input. The activation levels of the input units then propagate through connections to the output units, possibly mediated by one or more levels of intermediate units. The pattern of activation of the output units corresponds to the output of the computation and can be fed into a subsequent network or into response effectors. Many models of perceptual and cognitive processes within this family have been explored recently (for a recent collection of reports, including extensive tutorials, reviews, and historical surveys, see Rumelhart, McClelland, & The PDP Research Group, 1986; and McClelland, Rumelhart, & The PDP Research Group, 1986; henceforth, "PDPI" and "PDPII").

There is no doubt that these models have a different feel than standard symbol-processing models. The units, the topology and weights of the connections among them, the functions by which activation levels are transformed in units and connections, and the learning (i.e., weight-adjustment) function are all that is "in" these models; one cannot easily point to rules, algorithms, expressions, and the like inside them. By itself, of course, this means little, because the same is true for a circuit diagram of a digital computer implementing a theorem-prover. How, then, are PDP models related to the more traditional symbol-processing models that have until now dominated cognitive psychology and linguistics?

It is useful to distinguish three possibilities. In one, PDP models would occupy an intermediate level between symbol processing and neural hardware: they would characterize the elementary information processes provided by neural networks that serve as the building blocks of rules or algorithms. Individual PDP networks would compute the primitive symbol associations (such as matching an input against memory, or pairing the input and output of a rule), but the way the overall output of one network feeds into the input of another would be isomorphic to the structure of the symbol manipulations captured in the statements of rules. Progress in PDP modeling would undoubtedly force revisions in traditional models, because traditional assumptions about primitive mechanisms may be neurally implausible, and complex chains of symbol manipulations may be obviated by unanticipated primitive computational powers of PDP networks. Nonetheless, in this scenario a well-defined division between rule and hardware would remain, each playing an indispensable role in the explanation of a cognitive process. Many existing types of symbol-processing models would survive mostly intact, and, to the extent they have empirical support and explanatory power, would dictate many fundamental aspects of network organization. In some expositions of PDP models, this is the proposed scenario (see, e.g., Hinton, 1981; Hinton, McClelland, & Rumelhart, 1986, p. 78; also Touretzky, 1986 and Touretzky & Hinton, 1985, where PDP networks implement aspects of LISP and production systems, respectively). We call this "implementational connectionism".

An alternative possibility is that once PDP network models are fully developed, they will *replace* symbol-processing models as explanations of cognitive processes. It would be impossible to find a principled mapping between the components of a PDP model and the steps or memory structures implicated by a symbol-processing theory, to find states of the PDP model that correspond to intermediate states of the execution of the program, to observe stages of its growth corresponding to components of the program being put into place, or states of breakdown corresponding to components wiped out through trauma or loss—the structure of the symbolic model would vanish. Even the input–output function computed by the network model could differ in special cases from that computed by the symbolic model. Basically, the entire operation of the model (to the extent that it is not a black box) would have to be characterized not in terms of interactions among entities possessing *both* semantic *and* physical properties (e.g., different subsets of neurons or states of neurons each of which represent a distinct chunk of knowledge), but in terms of entities that had *only* physical properties, (e.g., the "energy landscape" defined by the activation levels of a large aggregate of interconnected neurons). Perhaps the symbolic model, as an approximate description of the performance in question, would continue to be useful as a heuristic, capturing some of the regularities in the domain in an intuitive or easily-communicated way, or allowing one to make convenient approximate predictions. But the symbolic model would not be a literal account at any level of analysis of what is going on in the brain, only an analogy or a rough summary of regularities. This scenario, which we will call "eliminative connectionism", sharply contrasts with the hardware–software distinction that has been assumed in cognitive science until now: no one would say that a program is an "approximate" description of the behavior of a computer, with the "exact" description existing at the level of chips and circuits; rather they are both exact descriptions at different levels of analysis.

Finally, there is a range of intermediate possibilities that we have already hinted at. A cognitive process might be profitably understood as a sequence or system of isolable entities that would be symbolic inasmuch as one could characterize them as having semantic properties such as truth values, consistency relations, or entailment relations, and one might predict the input–output function and systematicities in performance, development, or loss strictly in terms of formal properties of these entities. However, they might bear little resemblance to the symbolic structures that one would posit by studying a domain of intelligence independent of implementation considerations. The primitive information-processing operations made available by the connectionist architecture (summation of weighted activation levels and threshold functions, etc.) might force a theorist to posit a radically different set of symbols and operations, which in turn would make different predictions about the functions that could be computed and the patterns of breakdown observable during development, disease, or intermediate stages of processing. In this way, PDP theory could lead to fundamental new discoveries about the character of symbol-processing, rather than implying that there was no such thing. Let us call this intermediate position "revisionist-symbol-processing connectionism".

Language: A crucial test case. From its inception, the study of language within the framework of generative grammar has been a prototypical example of how fundamental properties of a cognitive domain can be explained within the symbolic paradigm. Linguistic theories have posited symbolic representations, operations, and architectures of rule-systems that are highly structured, detailed, and constrained, testing them against the plentiful and complex data of language (both the nature of adults' mastery of language, and data about how such knowledge is learned and put to use in comprehension and speech). Historically, it has been the demands of these theories that have driven our conception of what the computational resources underlying cognition must provide at a minimum (e.g., Chomsky, 1957, 1965). A priori notions of neurally possible elementary information processes have been plainly too weak, at worst, or unenlightening because of the few constraints they impose, at best. Language has been the domain most demanding of articulated symbol structures governed by rules and principles and it is also the domain where such structures have been explored in the greatest depth and sophistication, within a range of theoretical frameworks and architectures, attaining a wide variety of significant empirical results. Any alternative model that either eschews symbolic mechanisms altogether, or that is strongly shaped by the restrictive nature of available elementary information processes and unresponsive to the demands of the high-level functions being computed, starts off at a seeming disadvantage. Many observers thus feel that connectionism, as a radical restructuring of cognitive theory, will stand or fall depending on its ability to account for human language.

1.2. The Rumelhart–McClelland model and theory

One of the most influential efforts in the PDP school has been a model of the acquisition of the marking of the past tense in English developed by David Rumelhart and James McClelland (1986b, 1987). Using standard PDP mechanisms, this model learns to map representations of present tense forms of English verbs onto their past tense versions. It handles both regular (*walk/walked*) and irregular (*feel/felt*) verbs, productively yielding past forms for novel verbs not in its training set, and it distinguishes the variants of the past tense morpheme (*t* versus *d* versus *id*) conditioned by the final consonant of the verb (*walked* versus *jogged* versus *sweated*). Furthermore, in doing so it displays a number of behaviors reminiscent of children. It passes through stages of conservative acquisition of correct irregular and regular verbs (*walked, brought, hit*) followed by productive application of the regular rule and overregularization to irregular stems (e.g. *bringed, hitted*), followed by mastery of both regular and irregular verbs. It acquires subclasses of irregular verbs (e.g. *fly/flew, sing/sang, hit/hit*) in an order similar to children. It makes certain types of errors (*ated, wented*) at similar stages. Nonetheless, nothing in the model corresponds in any obvious way to the rules that have been assumed to be an essential part of the explanation of the past tense formation process. None of the individual units or connections in the model corresponds to a word, a position within a word, a morpheme, a regular rule, an exception, or a paradigm. The intelligence of the model is distributed in the pattern of weights linking the simple input and output units, so that any relation to a rule-based account is complex and indirect at best.

Rumelhart and McClelland take the results of this work as strong support for eliminative connectionism, the paradigm in which rule- or symbol-based accounts are simply eliminated from direct explanations of intelligence:

> We suggest instead that implicit knowledge of language may be stored in connections among simple processing units organized into networks. While the behavior of such networks may be describable (at least approximately) as conforming to some system of rules, we suggest that an account of the fine structure of the phenomena of language use and language acquisition can best be formulated in models that make reference to the characteristics of the underlying networks. (Rumelhart & McClelland, 1987, p. 196)

> 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. (Rumelhart & McClelland, 1986b, p. 267, their emphasis)

> 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. (Rumelhart & McClelland, 1986b, p. 268)

The Rumelhart–McClelland (henceforth, "RM") model, because it inspires these remarkable claims, figures prominently in general expositions of connectionism that stress its revolutionary nature, such as Smolensky (in press) and McClelland, Rumelhart, and Hinton (1986). Despite the radical nature of these conclusions, it is our impression that they have gained acceptance in many quarters; that many researchers have been persuaded that theories of language couched in terms of rules and rule acquisition may be obsolete (see, e.g., Sampson, 1987). Other researchers have attempted to blunt the force of the Rumelhart and McClelland's attack on rules by suggesting that the model really does contain rules, or that past tense acquisition is an unrepresentatively easy problem, or that there is some reason in principle why PDP models are incapable of being extended to language as a whole, or that Rumelhart and McClelland are modeling 'performance' and saying little about 'competence' or are modeling implementations but saying little about algorithms. We believe that these quick reactions—be they conversion experiences or outright dismissals—are unwarranted. Much can be gained by taking the model at face value as a theory of the psychology of the child and by examining the claims of the model in detail. That is the goal of this paper.

The RM model, like many PDP models, is a *tour de force*. It is explicit and mechanistic: precise empirical predictions flow out of the model as it operates autonomously, rather than being continuously molded or reshaped to fit the facts by a theorist acting as *deus ex machina*. The authors have made a commitment as to the underlying computational architecture of the model, rather than leaving it as a degree of freedom. The model is tested not only against the phenomenon that inspired it—the three-stage developmental sequence of generalizations of the regular past tense morpheme—but against several unrelated phenomena as well. Furthermore, Rumelhart and McClelland bring these developmental data to bear on the model in an unusually detailed way, examining not only gross effects but also many of its more subtle details. Several non-obvious but interesting empirical predictions are raised in these examinations. Finally, the model uses clever mechanisms that operate in surprising ways. These features are virtually unheard of in developmental psycholinguistics (see Pinker, 1979; Wexler & Culicover, 1980). There is no doubt that our understanding of language acquisition would advance more rapidly if theories in developmental psycholinguistics were held to such standards.

Nonetheless, our analysis of the model will come to conclusions very different from those of Rumelhart and McClelland. In their presentation, the model is evaluated only by a global comparison of its overall output behavior with that of children. There is no unpacking of its underlying theoretical assumptions so as to contrast them with those of a symbolic rule-based alternative, or indeed any alternative. As a result, there is no apportioning of credit or blame for the model's performance to properties that are essential versus accidental, or unique to it versus shared by any equally explicit alternative. In particular, Rumelhart and McClelland do not consider what it is about the standard symbol-processing theories that makes them "standard", beyond their first-order ability to relate stem and past tense. To these ends, we analyze the assumptions and consequences of the RM model, as compared to those of symbolic theories, and point out the crucial tests that distinguish them. In particular, we seek to determine whether the RM model is viable as a theory of human language acquisition—there is no question that it is a valuable demonstration of some of the surprising things that PDP models are capable of, but our concern is whether it is an accurate model of children.

Our analysis will lead to the following conclusions:

- Rumelhart and McClelland's actual explanation of children's stages of regularization of the past tense morpheme is demonstrably incorrect.
- Their explanation for one striking type of childhood speech error is also incorrect.
- Their other apparent successes in accounting for developmental phenomena either have nothing to do with the model's parallel distributed processing architecture, and can easily be duplicated by symbolic models, or involve major confounds and hence do not provide clear support for the model.
- The model is incapable of representing certain kinds of words.
- It is incapable of explaining patterns of psychological similarity among words.
- It easily models many kinds of rules that are not found in any human language.
- It fails to capture central generalizations about English sound patterns.
- It makes false predictions about derivational morphology, compounding, and novel words.
- It cannot handle the elementary problem of homophony.
- It makes errors in computing the past tense forms of a large percentage of the words it is tested on.
- It fails to generate any past tense form at all for certain words.
- It makes incorrect predictions about the reality of the distinction between regular rules and exceptions in children and in languages.

We will conclude that the claim that parallel distributed processing networks can eliminate the need for rules and for rule induction mechanisms in the explanation of human language is unwarranted. In particular, we argue that the shortcomings are in many cases due to central features of connectionist ideology and irremediable; or if remediable, only by copying tenets of the maligned symbolic theory. The implications for the promise of connectionism in explicating language are, we think, profound.

The paper is organized as follows. First, we examine in broad outline the phenomena of English verbal inflection. Then we describe the operation of the RM model and how it contrasts with the rule-based alternative, evaluating the merits of each. This amounts to an evaluation of the model in terms of its ability to handle the empirical properties of language in its adult state. In the next major section, we evaluate the model in terms of its ability to handle the empirical properties of children's path of development toward the adult state, comparing it with a simple model of symbolic rule acquisition. Finally, we evaluate the status of the radical claims about connectionism that were motivated by the RM model, and we determine the extent to which the performance of the RM model is a direct consequence of properties of its PDP architecture and thus bears on the promise of parallel distributed processing models in accounting for language and language acquisition.

2. A brief overview of English verbal inflection

2.1. *The basic facts of English inflection*

Rumelhart and McClelland aim to describe part of the system of verbal inflection in English. As background to our examination of their model, we briefly review the structure of the English verb, and present the basic flavor of a rule-based account of it.[1] When we evaluate the RM model, many additional details about the facts of English inflection and about linguistic theories of its structure will be presented.

English inflectional morphology is not notably complicated. Where the verb of classical Greek has about 350 distinct forms and the verb of current Spanish or Italian about 50, the regular English verb has exactly four:

[1] Valuable linguistic studies of the English verbal system include Bloch (1947), Bybee and Slobin (1982), Curme (1935), Fries (1940), Hoard and Sloat (1973), Hockett (1942), Jespersen (1942), Mencken (1936), Palmer (1930), Sloat and Hoard (1971), Sweet (1892). Chomsky and Halle (1968) and Kiparsky (1982a, b) are important general works touching on aspects of the system.

(1) a. walk
b. walks
c. walked
d. walking

As is typical in morphological systems, there is rampant syncretism—use of the same phonological form to express different, often unrelated morphological categories. On syntactic grounds we might distinguish 13 categories filled by the four forms.

(2) a. *-Ø* Present-everything but 3rd person singular:
I, you, we, they *open.*
Infinitive:
They may *open,* They tried to *open.*
Imperative:
Open!
Subjunctive:
They insisted that it *open.*

b. *-s* Present- 3rd person singular:
He, she, it *opens.*

c. *-ed* Past:
It *opened.*
Perfect Participle:
It has *opened.*
Passive Participle:
It was being *opened.*
Verbal adjective:
A recently-*opened* box.

d. *-ing* Progressive Participle:
He is *opening.*
Present Participle:
He tried *opening* the door.
Verbal noun (gerund):
His incessant *opening* of the boxes.
Verbal adjective:
A quietly-*opening* door.

The system is rendered more interesting by the presence of about 180 'strong' or 'irregular' verbs, which form the past tense other than by simple suffixation. There are, however, far fewer than 180 ways of modifying a stem to produce a strong past tense; the study upon which Rumelhart and McClelland depend, Bybee and Slobin (1982), divides the strong group into nine coarse and somewhat heterogeneous subclasses, which we discuss later. (See the Appendix for a précis of the entire system.)

Many strong verbs also maintain a further formal distinction, lost in (2c), between the past tense itself and the Perfect/Passive Participle, which is frequently marked with *-en*: 'he *ate*' vs. 'he has, was *eaten*'. These verbs mark the outermost boundary of systematic complexity in English, giving the learner five forms to keep track of, two of which—past and perfect/passive participle—are not predictable from totally general rules.[2]

2.2. Basic features of symbolic models of inflection

Rumelhart and McClelland write that "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." What they mean is that when Berko (1958) first documented children's ability to inflect novel verbs for past tense (e.g. *jicked*), and when Ervin (1964) documented overregularizations of irregular past tense forms in spontaneous speech (e.g. *breaked*), it was effective evidence against any notion that language acquisition consisted of rote imitation. But it is important to note the general point that the ability to generalize beyond rote forms is *not* the only motivation for using rules (as behaviorists were quick to point out in the 1960s when they offered their own accounts of generalization). In fact, even the existence of competing modes of generalizing, such as the different past tense forms of regular and irregular verbs or of regular verbs ending in different consonants, is not the most important motivation for positing distinct rules. Rather, rules are generally invoked in linguistic explanations in order to *factor* a complex phenomenon into simpler components that feed representations into one another. Different types of rules apply to these intermediate representations, forming a cascade of structures and rule components. Rules are individuated not only because they *compete* and mandate different transformations of the same input structure (such as *break—breaked/broke*), but because they apply to different *kinds* of structures, and thus impose a factoring of a phenomenon into distinct components, rather than generating the phenomena in a single step mapping inputs to outputs. Such factoring allows orthogonal generalizations to be extracted and stated separately, so that observed complexity can arise through interaction and feeding of independent rules and processes, which often have rather

[2] Somewhat beyond this bound lies the verb 'be' with eight distinct forms: *be, am, is, are, was, were, been, being*, of which only the last is regular.

different parameters and domains of relevance. This is immediately obvious in most of syntax, and indeed, in most domains of cognitive processing (which is why the acquisition and use of internal representations in "hidden units" is an important technical problem in connectionist modeling; see Hinton & Sejnowski, 1986; Rumelhart, Hinton, & Williams, 1986).

However, it is not as obvious at first glance how rules feed each other in the case of past tense inflection. Thus to examine in what sense the RM model "has no rules" and thus differs from symbolic accounts, it is crucial to spell out how the different rules in the symbolic accounts are individuated in terms of the components they are associated with.

There is one set of "rules" inherent in the generation of the past tense in English that is completely outside the mapping that the RM model computes: those governing the interaction between the use of the past tense form and the type of sentence the verb appears in, which depends on semantic factors such as the relationship between the times of the speech act, referent event, and a reference point, combined with various syntactic and lexical factors such as the choice of a matrix verb in a complex sentence (*I helped her leave/*left* versus *I know she left/*leave*) and the modality and mood of a sentence (*I went/*go yesterday* versus *I didn't go/*went yesterday; If my grandmother had/*has balls she'd be my grandfather*). In other words, a speaker doesn't choose to produce a past tense form of a verb when and only when he or she is referring to an event taking place before the act of speaking. The distinction between the mechanisms governing these phenomena, and those that associate individual stems and past tense forms, is implicitly accepted by Rumelhart and McClelland. That is, presumably the RM model would be embedded in a collection of networks that would pretty much reproduce the traditional picture of there being one set of syntactic and semantic mechanisms that selects occasions for use of the past tense, feeding information into a distinct morphological-phonological system that associates individual stems with their tensed forms. As such, one must be cautious at the outset in saying that the RM model is an alternative to a rule-based account of the past tense in general; at most, it is an alternative to whatever decomposition is traditionally assumed within the part of grammar that associates stems and past tense forms.[3]

In symbolic accounts, this morphological-phonological part is subject to

[3] Furthermore, the RM model seeks only to *generate* past forms from stems; it has no facility for retrieving a stem given the past tense form as input. (There is no guarantee that a network will run 'backwards' and in fact some of the more sophisticated learning algorithms presuppose a strictly feed-forward design.) Presumably the human learner can go both ways from the very beginning of the process; later we present examples of children's back-formations in support of this notion. Rule-based theories, as accounts of knowledge rather than use of knowledge, are neutral with respect to the production/recognition distinction.

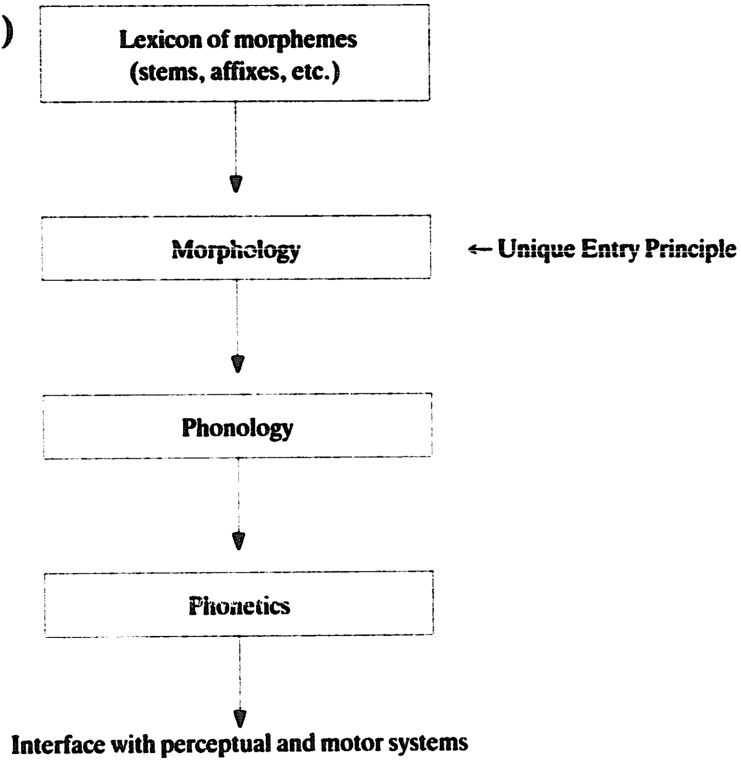
further decomposition. In particular, rule-based accounts rely on several fundamental distinctions:

- *Lexical item vs. phoneme string*. The lexical item is a unique, idiosyncratic set of syntactic, semantic, morphological, and phonological properties. The phoneme string is just one of these properties. Distinct items may share the same phonological composition (homophony). Thus the notion of lexical representation distinguishes phonologically ambiguous words such as *wring* and *ring*.
- *Morphological category vs. morpheme*. There is a distinction between a morphological category, such as 'past tense' or 'perfect aspect' or 'plural' or 'nominative case', and the realization(s) of it in phonological substance. The relation can be many-one in both directions: the same phonological entity can mark several categories (syncretism); and one category may have several (or indeed many) realizations, such as through a variety of suffixes or through other means of marking. Thus in English, *-ed* syncretistically marks the past, the perfect participle, the passive participle, and a verbal adjective—distinct categories; while the past tense category itself is manifested differently in such items as *bought, blew, sat, bled, bent, cut, went, ate, killed*.
- *Morphology vs. phonology*. Morphological rules describe the syntax of words—how words are built from morphemes—and the realization of abstract morphological categories. Phonological rules deal with the predictable features of sound structure, including adjustments and accommodations occasioned by juxtaposition and superposition of phonological elements. Morphology trades in such notions as 'stem', 'prefix', 'suffix', 'past tense'; phonology in such as 'vowel', 'voicing', 'obstruence', 'syllable'. As we will see in our examination of English morphology, there can be a remarkable degree of segregation of the two vocabularies into distinct rule systems: there are morphological rules which are blind to phonology, and phonological rules blind to morphological category.
- *Phonology vs. phonetics*. Recent work (Liberman & Pierrehumbert, 1984; Pierrehumbert & Beckman, 1986) refines the distinction between phonology proper, which establishes and maps between one phonological representation and another, and phonetic implementation, which takes a representation and relates it to an entirely different system of parameters (for example, targets in acoustic or articulatory space).

In addition, a rule-system is organized by principles which determine the interactions between rules: whether they compete or feed, and if they compete, which wins. A major factor in regulating the feeding relation is organization into components: morphology, an entire set of formation rules, feeds phonology, which feeds phonetics.[4] Competition among morphological alternatives is under the control of a principle of paradigm structure (called the 'Unique Entry Principle' in Pinker, 1984) which guarantees that in general each word will have one and only one form for each relevant morphological category; this is closely related to the 'Elsewhere Condition' of formal linguistics (Kiparsky, 1982a, b). The effect is that when a general rule (like Past(x) = x + *ed*) formally overlaps a specific rule (like Past(*go*) = *went*), the specific rule not only applies but also blocks the general one from applying.

The picture that emerges looks like this:

(3)


Lexicon of morphemes (stems, affixes, etc.)
Morphology
← Unique Entry Principle
Phonology
Phonetics
Interface with perceptual and motor systems

With this general structure in mind, we can now examine how the RM model differs in "not having rules".

3. The Rumelhart–McClelland model

Rumelhart and McClelland's goal is to model the acquisition of the past tense, specifically the *production* of the past tense, considered in isolation

[4] More intricate variations on this basic pattern are explored in recent work in "Lexical Phonology"; see Kiparsky (1982a, b).

from the rest of the English morphological system. They assume that the acquisition process establishes a direct mapping from the phonetic representation of the stem to the phonetic representation of the past tense form. The model therefore takes the following basic shape:

(4) Uninflected stem → Pattern associator → Past form

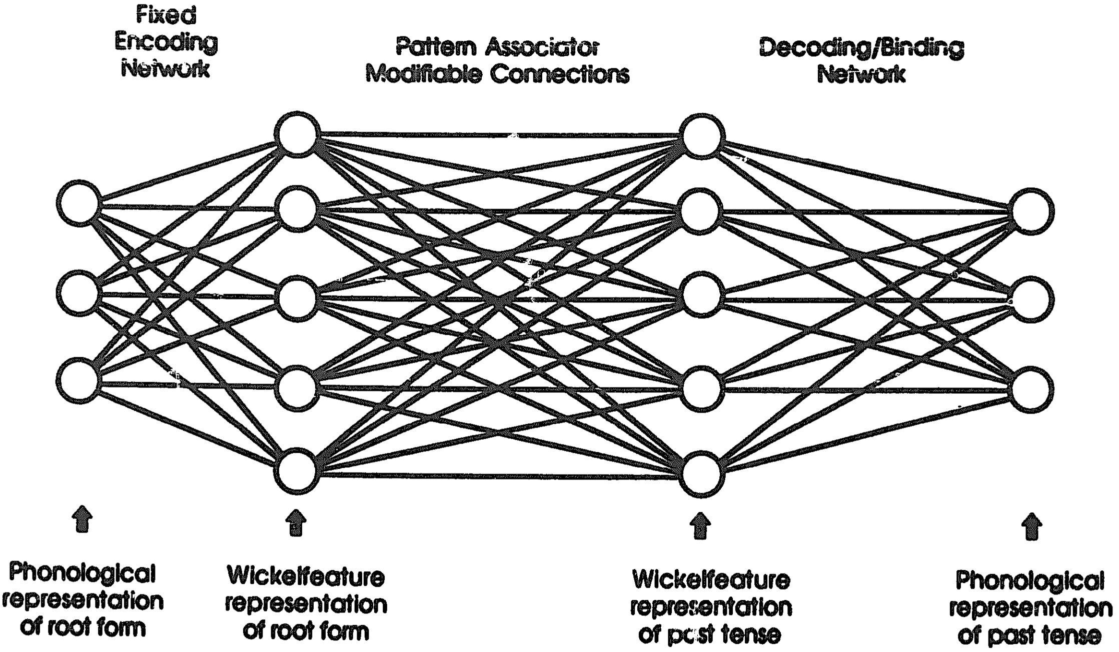
This proposed organization of knowledge collapses the major distinctions embodied in the linguistic theory sketched in (3). In the following sections we ascertain and evaluate the consequences of this move.

The detailed structure of the RM model is portrayed in Figure 1.

In its trained state, the pattern associator is supposed to take any stem as input and emit the corresponding past tense form. The model's pattern associator is a simple network with two layers of nodes, one for representing input, the other for output. Each node represents a different property that an input item may have. Nodes in the RM model may only be 'cn' or 'off'; thus the nodes represent binary features, 'off' and 'on' marking the simple absence or presence of a certain property. Each stem must be encoded as a unique subset of turned-on input nodes; each possible past tense form as a unique subset of output nodes turned on.

Here a nonobvious problem asserts itself. The natural assumption would be that words are strings on an alphabet, a concatenation of phonemes. But

Figure 1. *The Rumelhart-McClelland model of past tense acquisition. (Reproduced from Rumelhart and McClelland, 1986b, p. 222, with permissic: of the publisher, Bradford Books/MIT Press.)*


Fixed Encoding Network
Pattern Associator Modifiable Connections
Decoding/Binding Network
Phonological representation of root form
Wickelfeature representation of root form
Wickelfeature representation of past tense
Phonological representation of past tense

each datum fed to a network must decompose into an unordered set of properties (coded as turned-on units), and a string is a prime example of an *ordered* entity. To overcome this, Rumelhart and McClelland turn to a scheme proposed by Wickelgren (1969), according to which a string is represented as the set of the trigrams (3-character-sequences) that it contains. (In order to locate word-edges, which are essential to phonology and morphology, it is necessary to assume that 'word-boundary' (#) is a character in the underlying alphabet.) Rumelhart and McClelland call such trigrams *Wickelphones*. Thus a word like *strip* translates, in their notation to {$_\#s_t$, $_st_r$, $_tr_i$, $_ri_p$, $_ip_\#$}. Note that the word *strip* is uniquely reconstructible from the cited trigram set. Although certain trigram sets are consistent in principle with more than one string, Rumelhart and McClelland find that all words in their sample are uniquely encoded. Crucially, each possible trigram must be construed as an atomic property that a string may have or lack. Thus, writing it out as we did above is misleading, because the order of the five Wickelphones is not represented anywhere in the RM system, and there is no selective access to the "central" phoneme *t* in a Wickelphone $_st_r$ or to the "context" phonemes $_sX$ and X_r. It is more faithful to the actual mechanism to list the Wickelphones in arbitrary (e.g., alphabetical) order and avoid any spurious internal decomposition of Wickelphones, hence: {*ip#, rip, str, tri, #st*}.

For immediate expository purposes, we can think of each unit in the input layer of the networks as standing for one of the possible Wickelphones; likewise for each unit in the output layer. Any given word is encoded as a pattern of node activations over the whole set of Wickelphone nodes—as a set of Wickelphones. This gives a "distributed" representation: an individual word does not register on its own node, but is analyzed as an ensemble of properties, Wickelphones, which are the true primitives of the system. As Figure 1 shows, Rumelhart and McClelland require an "encoder" of unspecified nature to convert an ordered phonetic string into a set of activated Wickelphone units; we discuss some of its properties later.

The Wickelphone contains enough context to detect in gross the kind of input–output relationships found in the stem-to-past tense mapping. Imagine a pattern associator mapping from input Wickelphones to output Wickelphones. As is usual in such networks, every input node is connected to every output node, giving each input Wickelphone node the chance to influence every node in the output Wickelphone set. Suppose that a set of input nodes is turned on, representing an input to the network. Whether a given output node will turn on is determined jointly by the strength of its connections to the active input nodes and by the output node's own overall susceptibility to influence, its 'threshold'. The individual on/off decisions for the output units are made probabilistically, on the basis of the discrepancy between total input and threshold: the nearer the input is to the threshold, the more random the decision.

An untrained pattern associator starts out with no preset relations between input and output nodes—link weights at zero—or with random input–output relations; it's a tabula rasa that is either blank or meaninglessly noisy. (Rumelhart & McClelland's is blank.) Training involves presenting the network with an input form (in the present case, a representation of a stem) and comparing the output pattern actually obtained with the desired pattern for the past tense form, which is provided to the network by a "teacher" as a distinct kind of "teaching" input (not shown in Figure 1). The corresponding psychological assumption is that the child, through some unspecified process, has already figured out which past tense form is to be associated with which stem form. We call this the "juxtaposition process"; Rumelhart and McClelland adopt the not unreasonable idealization that it does not interact with the process of abstracting the nature of the mapping between stem and past forms.

The comparison between the actual output pattern computed by the connections between input and output nodes, and the desired pattern provided by the "teacher", is made on a node-by-node basis. Any output node that is in the wrong state becomes the target of adjustment. If the network ends up leaving a node off that ought to be on according to the teacher, changes are made to render that node more likely to fire in the presence of the particular input at hand. Specifically, the weights on the links connecting active input units to the recalcitrant output unit are increased slightly; this will increase the tendency for the currently active input units—those that represent the input form—to activate the target node. In addition, the target node's own threshold is lowered slightly, so that it will tend to turn on more easily across the board. If, on the other hand, the network incorrectly turns an output node *on*, the reverse procedure is employed: the weights of the connections from currently active input units are decremented (potentially driving the connection weight to a negative, inhibitory value) and the target node's threshold is raised; a hyperactive output node is thus made more likely to turn off given the same pattern of input node activation. Repeated cycling through input–output pairs, with concomitant adjustments, shapes the behavior of the pattern associator. This is the "perceptron convergence procedure" (Rosenblatt, 1962) and it is known to produce, in the limit, a set of weights that successfully maps the input activation vectors onto the desired output activation vectors, as long as such a set of weights exists.

In fact, the RM net, following about 200 training cycles of 420 stem-past pairs (a total of about 80,000 trials), is able to produce correct past forms for the stems when the stems are presented alone, that is, in the absence of "teaching" inputs. Somewhat surprisingly, a single set of connection weights in the network is able to map *look* to *looked*, *live* to *lived*, *melt* to *melted*, *hit* to *hit*, *make* to *made*, *sing* to *sang*, even *go* to *went*. The bits of stored information accomplishing these mappings are superimposed in the connection weights and node thresholds; no single parameter corresponds uniquely to a rule or to any single irregular stem-past pair.

Of course, it is necessary to show how such a network generalizes to stems it has not been trained on, not only how it reproduces a rote list of pairs. The circumstances under which generalization occurs in pattern associators with distributed representations is reasonably well understood. Any encoded (one is tempted to say 'en-noded') property of the input data that participates in a frequently attested pattern of input/output relations will play a major role in the development of the network. Because it is turned on during many training episodes, and because it stands in a recurrent relationship to a set of output nodes, its influence will be repeatedly enhanced by the learning procedure. A connectionist network does more than match input to output; it responds to regularities in the representation of the data and uses them to accomplish the mapping it is trained on and to generalize to new cases. In fact, the distinction between reproducing the memorized input–output pairs and generating novel outputs for novel inputs is absent from pattern associators: a single set of weights both reproduces trained pairs and produces novel outputs which are blends of the output patterns strongly associated with each of the properties defining the novel input.

The crucial step is therefore the first one: coding the data. If the patterns in the data relevant to generalizing to new forms are not encoded in the representation of the data, no network—in fact, no algorithmic system of any sort—will be able to find them. (This is after all the reason that so much research in the 'symbolic paradigm' has centered on the nature of linguistic representations.) Since phonological processes and relations (like those involved in past tense formation) do not treat phonemes as atomic, unanalyzable wholes but refer instead to their constituent phonetic properties like voicing, obstruency, tenseness of vowels, and so on, it is necessary that such fine-grained information be present in the network. The Wickelphone, like the phoneme, is too coarse to support generalization. To take an extreme example adapted from Morris Halle, any English speaker who labors to pronounce the celebrated composer's name as [bax] knows that if there were a verb *to Bach*, its past would be *baxt* and not *baxd* or *baxid*, even though no existing English word contains the velar fricative [x]. Any representation that does not characterize *Bach* as similar to *pass* and *walk* by virtue of ending in an unvoiced segment would fail to make this generalization. Wickelphones, of course, have this problem; they treat segments as opaque quarks and fail to display vital information about segmental similarity classes. A better representation would have units referring in some way to phonetic features rather than to phonemes, because of the well-known fact that the correct dimension of generalization from old to new forms must be in terms of such features.

Rumelhart and McClelland present a second reason for avoiding Wickelphone nodes. The number of possible Wickelphones for their representation of English is $35^3 + (2 \times 35^2) = 45{,}325$ (all triliterals + all biliterals beginning and ending with #). The number of distinct connections from the entire input Wickelvector to its output clone would be over two billion ($45{,}325^2$), too many to handle comfortably. Rumelhart and McClelland therefore assume a phonetic decomposition of segments into features which are in broad outline like those of modern phonology. On the basis of this phonetic analysis, a Wickelphone dissolves into a set of 'Wickelfeatures', a sequence of three features, one from each of the three elements of the Wickelphone. For example, the features "VowelUnvoicedInterrupted" and "HighStopStop" are two of the Wickelfeatures in the ensemble that would correspond to the Wickelphone "ipt". In the RM model, units represent Wickelfeatures, not Wickelphones; Wickelphones themselves play no role in the model and are only represented implicitly as sets of Wickelfeatures. Again, there is the potential for nondistinct representations, but it never occurred in practice for their verb set. Notice that the actual atomic properties recognized by the model are not phonetic features per se, but entities that can be thought of as 3-feature sequences. The Wickelphone/Wickelfeature is an excellent example of the kind of novel properties that revisionist-symbol-processing connectionism can come up with.

A further refinement is that not all definable Wickelfeatures have units dedicated to them: the Wickelfeature set was trimmed to exclude, roughly, feature-triplets whose first and third features were chosen from different phonetic dimensions.[5] The end result is a system of 460 nodes, each one representing a Wickelfeature. One may calculate that this gives rise to $460^2 = 211{,}600$ input–output connections.

The module that encodes words into input Wickelfeatures (the "Fixed Encoding Network" of Figure 1) and the one that decodes output Wickelfeatures into words (the "Decoding/Binding Network" of Figure 1) are perhaps not meant to be taken entirely seriously in the current implementation of the RM model, but several of their properties are crucially important in under-

[5] Although this move was inspired purely by considerations of computational economy, it or something like it has real empirical support; the reader familiar with current phonology will recognize its relation to the notion of a 'tier' of related features in autosegmental phonology.

standing and evaluating it. The input encoder is deliberately designed to activate some incorrect Wickelfeatures in addition to the precise set of Wickelfeatures in the stem: specifically, a randomly selected subset of those Wickelfeatures that encode the features of the central phoneme properly but encode incorrect feature values for one of the two context phonemes. This "blurred" Wickelfeature representation cannot be construed as random noise; the same set of incorrect Wickelfeatures is activated every time a word is presented, and no Wickelfeature encoding an incorrect choice of the *central* feature is ever activated. Rather, the blurred representation fosters generalization. Connectionist pattern associators are always in danger of capitalizing too much on idiosyncratic properties of words in the training set in developing their mapping from input to output and hence of not properly generalizing to new forms. Blurring the input representations makes the connection weights in the RM model less likely to be able to exploit the idiosyncrasies of the words in the training set and hence reduces the model's tendency toward conservatism.

The output decoder faces a formidable task. When an input stem is fed into the model, the result is a set of activated output Wickelfeature units. Which units are on in the output depends on the current weights of the connections from active input units and on the probabilistic process that converts the summed weighted inputs into a decision as to whether or not to turn on. Nothing in the model ensures that the set of activated output units will fit together to describe a legitimate word: the set of activated units do not have to have neighboring context features that "mesh" and hence implicitly "assemble" the Wickelfeatures into a coherent string; they do not have to be mutually consistent in the feature they mandate for a given position; and they do not have to define a set of features for a given position that collectively define an English phoneme (or any kind of phoneme). In fact, the output Wickelfeatures virtually *never* define a word exactly, and so there is no clear sense in which one knows which word the output Wickelfeatures are defining. In many cases, Rumelhart and McClelland are only interested in assessing how likely the model seems to be to output a given target word, such as the correct past tense form for a given stem; in that case they can peer into the model, count the number of desired Wickelfeatures that are successfully activated and vice versa, and calculate the goodness of the match. However, this does not reveal which phonemes, or which words, the model would actually output.

To assess how likely the model actually is to output a phoneme in a given context, that is, how likely a given *Wickelphone* is in the output, a *Wickelphone Binding Network* was constructed as part of the output decoder. This network has units corresponding to Wickelphones; these units "compete" with one another in an iterative process to "claim" the activated Wickelfeatures: the more Wickelfeatures that a Wickelphone unit uniquely accounts for, the greater its strength (Wickelfeatures accounted for by more than one Wickelphone are "split" in proportion to the number of other Wickelfeatures each Wickelphone accounts for uniquely) and, supposedly, the more likely that Wickelphone is to appear in the output. A similar mechanism, called the *Whole-String Binding Network*, is defined to estimate the model's relative tendencies to output any of a particular set of words when it is of interest to compare those words with one another as possible outputs. Rumelhart and McClelland choose a set of plausible output words for a given input stem, such as *break, broke, breaked* and *broked* for the past tense of *break*, and define a unit for each one. The units then compete for activated Wickelfeatures in the output vector, each one growing in strength as a function of the number of activated Wickelfeatures it uniquely accounts for (with credit for nonunique Wickelfeatures split between the words that can account for it), and diminishing as a function of the number of activated Wickelfeatures that are inconsistent with it. This amounts to a forced-choice procedure and still does not reveal what the model would output if left to its own devices—which is crucial in evaluating the model's ability to produce correct past tense forms for stems it has not been trained on. Rumelhart and McClelland envision an eventual "sequential readout process" that would convert Wickelfeatures into a single temporally ordered representation, but for now they make do with a more easily implemented substitute: an *Unconstrained Whole-String Binding Network*, which is a whole-string binding network with one unit for every possible string of phonemes less than 20 phonemes long—that is, a forced-choice procedure among all possible strings. Since this process would be intractable to compute on today's computers, and maybe even tomorrow's, they created whole-string units only for a sharply restricted subset of the possible strings, those whose Wickelphones exceed a threshold in the Wickelphone binding network competition. But the set was still fairly large and thus the model was in principle capable of selecting both correct past tense forms and various kinds of distortions of them. Even with the restricted set of whole strings available in the unconstrained whole-string binding network, the iterative competition process was quite time-consuming in the implementation, and thus Rumelhart and McClelland ran this network only in assessing the model's ability to produce past forms for untrained stems; in all other cases, they either counted features in the output Wickelfeature vector directly, or set up a restricted forced-choice test among a small set of likely alternatives in the whole-string binding network.

In sum, the RM model works as follows. The phonological string is cashed in for a set of Wickelfeatures by an unspecified process that activates all the correct and some of the incorrect Wickelfeature units. The pattern associator excites the Wickelfeature units in the output; during the training phase its parameters (weights and thresholds) are adjusted to reduce the discrepancy between the excited Wickelfeature units and the desired ones provided by the teacher. The activated Wickelfeature units may then be decoded into a string of Wickelphones by the Wickelphone binding network, or into one of a small set of words by the whole-string binding network, or into a free choice of an output word by the unconstrained whole-string binding network.

4. An analysis of the assumptions of the Rumelhart–McClelland model in comparison with symbolic accounts

It is possible to practice psycholinguistics with minimal commitment to explicating the internal representation of language achieved by the learner. Rumelhart and McClelland's work is emphatically not of this sort. Their model is offered precisely as a model of internal representation; the learning process is understood in terms of changes in a representational system as it converges on the mature state. It embodies claims of the greatest psycholinguistic interest: it has a theory of phonological representation, a theory of morphology, a theory (or rather anti-theory) of the role of the notion 'lexical item', and a theory of the relation between regular and irregular forms. In no case are these presupposed theories simply transcribed from familiar views; they constitute a bold new perspective on the central issues in the study of word-forms, rooted in the exigencies and strengths of connectionism.

The model largely exemplifies what we have called revisionist-symbol-processing connectionism, rather than implementational or eliminative connectionism. Standard symbolic rules are not embodied in it; nor does it posit an utterly opaque device whose operation cannot be understood in terms of symbol-processing of any sort. It is possible to isolate an abstract but unorthodox linguistic theory implicit in the model (though Rumelhart and McClelland do not themselves consider it in this light), and that theory can be analyzed and evaluated in the same way that more familiar theories are. These are the fundamental linguistic assumptions of the RM model:

- That the Wickelphone/Wickelfeature provides an adequate basis for phonological generalization, circumventing the need to deal with strings.
- That the past tense is formed by direct modification of the phonetics of the root, so that there is no need to recognize a more abstract level of morphological structure.
- That the formation of strong (irregular) pasts is determined by purely

phonetic considerations, so that there is no need to recognize the notion 'lexical item' to serve as a locus of idiosyncrasy.

- That the regular system is qualitatively the same as the irregular, differing only in the number and uniformity of their populations of exemplars, so that it is appropriate to handle the whole stem/past relation in a single, indissoluble facility.

These rather specific assumptions combine to support the broader claim that connectionism supplies a viable alternative to highly structured symbol-processing theories such as that sketched above. "We have shown," they write (PDPII, p. 267), "that a reasonable account of the acquisition of the past tense can be provided without recourse to the notion of a 'rule' as anything more than a description of the language." By this they mean that rules, as mere summaries of the data, are not intrinsically or causally involved in internal representations. Rumelhart and McClelland's argument for the broader claim is based entirely on the behavior of their model.

We will show that each of the listed assumptions grossly mischaracterizes the domain it is relevant to, in a way that seriously undermines the model's claim to accuracy and even 'reasonableness'. More positively, we will show how past tense formation takes its place within a larger, more inclusive system of phonological and morphological interactions. The properties of the larger system will provide us with a clear benchmark for measuring the value of linguistic and psycholinguistic models.

4.1. Wickelphonology

The Wickelphone/Wickelfeature has some useful properties. Rumelhart and McClelland hold that the finite Wickelphone set can encode strings of arbitrary length (PDPII, p. 269) and though false this is close enough to being true to give them a way to distinguish all the words in their data. In addition, a Wickelphone contains a chunk of context within which phonological dependencies can be found. These properties allow the RM model to get off the ground. If, however, the Wickelphone/Wickelfeature is to be taken seriously as even an approximate model of phonological representation, it must satisfy certain basic, uncontroversial criteria.[6]

Preserving distinctions. First of all, a phonological representation system for a language must preserve all the distinctions that are actually present in

[6] For other critiques of the Wickelphone hypothesis, antedating the RM model, see Halwes and Jenkins (1971) and Savin and Bever (1970).

the language. English orthography is a familiar representational system that fails to preserve distinctness: for example, the word spelled 'read' may be read as either [rid] or [rεd];[7] whatever its other virtues, spelling is not an appropriate medium for phonological computation. The Wickelphone system fails more seriously, because there are distinctions that it is in principle incapable of handling. Certain patterns of repetitions will map distinct string-regions onto the same Wickelphone set, resulting in irrecoverable loss of information. This is not just a mathematical curiosity. For example, the Australian language Oykangand (Sommer, 1980) distinguishes between *algal* 'straight' and *algalgal* 'ramrod straight', different strings which share the Wickelphone set {alg, al#, gal, lga, #al}, as can be seen from the analysis in (5):

(5) a. algal
#al
alg
lga
gal
al#

b. algalgal
#al
alg
lga
gal
alg
lga
gal
al#

Wickelphone sets containing subsets closed under cyclic permutation on the character string—{alg, gal, lga} in the example at hand—are infinitely ambiguous as to the strings they encode. This shows that Wickelphones cannot represent even relatively short strings, much less strings of arbitrary length, without loss of concatenation structure (loss is guaranteed for strings

[7] We will use the following phonetic notation and terminology (sparingly). Enclosure in square brackets [ ] indicates phonetic spelling.

The *tense* vowels are: [i] as in *beat* [u] as in *shoe*
[e] as in *bait* [o] as in *go*

The *lax* vowels are: [I] as in *bit* [U] as in *put*
[ε] as in *bet* [ɔ] as in *lost*

The low front vowel [æ] appears in *cat*. The low central vowel [ʌ] appears in *shut*. The low back vowel [ɔ] appears in *caught*. The diphthong [ay] appears in *might* and *bite*; the diphthong [aw] in *house*. The high lax central vowel [ɨ] is the second vowel in *melted, rose's*.
The symbol [č] stands for the voiceless palato-alveolar affricate that appears twice in *church*; the symbol [ǰ] for its voiced counterpart, which appears twice in *judge*. [š] is the voiceless palato-alveolar fricative of *shoe* and [ž] is its voiced counterpart, the final consonant of *rouge*. The velar nasal ŋ is the final consonant in *sing*.
The term *sonorant consonant* refers to the liquids *l,r* and the nasals *m,n,*ŋ. The term *obstruent* refers to the complement set of oral stops, fricatives and affricates, such as *p,t,k,f,s,š,č,b,d,g,v,z,ž,ǰ*. The term *coronal* refers to sounds made at the dental, alveolar, and palato-alveolar places of articulation. The term *sibilant* refers to the conspicuously noisy fricatives and affricates [s,z,š,ž,č,ǰ].

over a certain length). On elementary grounds, then, the Wickelphone is demonstrably inadequate.

Supporting generalizations. A second, more sophisticated requirement is that a representation supply the basis for proper generalization. It is here that the phonetic vagaries of the most commonly encountered representation of English—its spelling—receive a modicum of justification. The letter *i*, for example, is implicated in the spelling of both [ay] and [I], allowing word-relatedness to be overtly expressed as identity of spelling in many pairs like those in (6):

(6) a. wr*i*te-wr*i*tten
b. b*i*te-b*i*t
c. ign*i*te-ign*i*tion
d. sen*i*le-sen*i*lity
e. der*i*ve-der*i*vative

The Wickelphone/Wickelfeature provides surprisingly little help in finding phonological generalizations. There are two domains in which significant similarities are operative: (1) among items in the input set, and (2) between an input item and its output form. Taking the trigram as the primitive unit of description impedes the discovery of inter-item similarity relations.

Consider the fact, noted by Rumelhart and McClelland, that the word *silt* and the word *slit* have no Wickelphones in common: the first goes to {#si, sil, ilt, lt#}, the second to {#sl, sli, lit, it#}. The implicit claim is that such pairs have no phonological properties in common. Although this result meets the need to distinguish the distinct, it shows that Wickelphone composition is a very unsatisfactory measure of psychological phonetic similarity. Indeed, historical changes of the type *slit* → *silt* and *silt* → *slit*, based on phonetic similarity, are fairly common in natural language. In the history of English, for example, we find *hross* → *horse, thrid* → *third, brid* → *bird* (Jespersen, 1942, p. 58). On pure Wickelphones such changes are equivalent to complete replacements; they are therefore no more likely, and no easier to master, than any other complete replacement, like *horse* going to *slit* or *bird* to *clam*. The situation is improved somewhat by the transition to Wickelfeatures, but remains unsatisfactory. Since phonemes *l* and *i* share features like voicing, Wickelphones like *sil* and *sli* will share Wickelfeatures like Voiceless-Voiced-Voiced. The problem is that the *l/i* overlap is the same as the overlap of *l* with any vowel and the same as the overlap of *r* with vowels. In Wickelfeatures it is just as costly—counting by number of replacements—to turn *brid* to phonetically distant *bald* or *blud* as it is to turn it to nearby *bird*.

Even in the home territory of the past tense, Wickelphonology is more an encumbrance than a guide. The dominant regularity of the language entails that a verb like *kill* will simply add one phone [d] in the past; in Wickelphones the map is as in (7):

(7) a. {#ki, kil, il#} → {#ki, kil, ild, ld#}
b. il# → ild, ld#

The change, shown in (7b), is exactly the full replacement of one Wickelphone by two others. The Wickelphone is in principle incapable of representing an observation like 'add [d] to the end of a word when it ends in a voiced consonant', because there is no way to single out the one word-ending consonant and no way to add a phoneme without disrupting the stem; you must refer to the entire sequence AB#, whether A is relevant or not, and you must replace it entirely, regardless of whether the change preserves input string structure. Given time and space, the facts can be registered on a Wickelfeature-by-Wickelfeature basis, but the unifying pattern is undiscoverable. Since the relevant phonological process involves only a *pair* of representationally adjacent elements, the triune Wickelphone/Wickelfeature is quite generally incompetent to locate the relevant factors and to capitalize on them in learning, with consequences we will see when we examine the model's success in generalizing to new forms.

The "blurring" of the Wickelfeature representation, by which certain input units *X*BC and AB*Z* are turned on in addition to authentic ABC, is a tactical response to the problem of finding similarities among the input set. The reason that A*Y*B is not also turned on—as one would expect, if "blurring" corresponded to neural noise of some sort—is in part that XBC and ABZ are units preserving the empirically significant adjacency pairing of segments: in many strings of the form ABC, we expect interactions within AB and BC, but not between A and C. Blurring both A and C helps to model processes in which only the presence of B is significant, and as Lachter and Bever (1988) show, partially recreates the notion of the single phoneme as a phonological unit. Such selective "blurring" is not motivated within Rumelhart and McClelland's theory or by general principles of PDP architecture; it is an external imposition that pushes it along more or less in the right direction. Taken literally, it is scarcely credible: the idea would be that the pervasive adjacency requirement in phonological processes is due to quasi-random confusion, rather than structural features of the representational apparatus and the physical system it serves.

Excluding the impossible. The third and most challenging requirement we can place on a representational system is that it should exclude the impossible. Many kinds of formally simple relations are absent from natural language, presumably because they cannot be mentally represented. Here the Wickelphone/Wickelfeature fails spectacularly. A quintessential unlinguistic map is relating a string to its mirror image reversal (this would relate *pit* to *tip, brag* to *garb, dumb* to *mud*, and so on); although neither physiology nor physics forbids it, no language uses such a pattern. But it is as easy to represent and learn in the RM pattern associator as the identity map. The rule is simply to replace each Wickelfeature ABC by the Wickelfeature CBA. In network terms, assuming link-weights from 0 to 1, weight the lines from ABC → CBA at 1 and all the (459) others emanating from ABC at 0. Since all weights start at 0 for Rumelhart and McClelland, this is exactly as easy to achieve as weighting the lines ABC → ABC at 1, with the others from ABC staying at 0; and it requires considerably less modification of weights than most other input–output transforms. Unlike other, more random replacements, the $S \rightarrow S^R$ map is guaranteed to preserve the stringhood of the input Wickelphone set. It is easy to define other processes over the Wickelphone that are equally unlikely to make their appearance in natural language: for example, no process turns on the identity of the entire first Wickelphone (#AB) or last Wickelphone (AB#)—compare in this regard the notions 'first (last) segment', 'first (last) syllable', frequently involved in actual morphological and phonological processes, but which appear as arbitrary disjunctions, if reconstructible at all, in the Wickelphone representation. The Wickelphone tells us as little about unnatural avenues of generalization as it does about the natural ones.

The root cause, we suggest, is that the Wickelphone is being asked to carry two contradictory burdens. Division into Wickelphones is primarily a way of multiplying out possible rule-contexts in advance. Since many phonological interactions are segmentally local, a Wickelphone-like decomposition into short substrings will pick out domains in which interaction is likely.[8] But any such decomposition must also retain enough information to allow the string to be reconstituted with a fair degree of certainty. Therefore, the minimum usable unit to reconstruct order is three segments long, even though many contexts for actual phonological processes span a window of only two seg-

[8] Of course, not all interactions are segmentally local. In vowel harmony, for example, a vowel typically reacts to a nearby vowel over an intervening string of consonants; if there are two intervening consonants, the interacting vowels will never be in the same Wickelphone and generalization will be impossible. Stress rules commonly skip over a string of one or two *syllables*, which may contain many segments: crucial notions such as 'second syllable' will have absolutely no characterization in Wickelphonology (see Sietsema, 1987, for further discussion). Phenomena like these show the need for more sophisticated representational resources. so that the relevant notion of domain of interaction may be adequately defined (see van der Hulst & Smith, 1982, for an overview of recent work). It is highly doubtful that Wickelphonology can be strengthened to deal with such cases, but we will not explore these broader problems, because our goal is to examine the Wickelphone as an alternative to the segmental concatenative structure which every theory of phonology includes.

ments. Similarly, if the blurring process were done thoroughly, so that ABC would set off all XBZ in the input set, there would be a full representation of the presence of B, but the identity of the input string would disappear. The RM model thus establishes a mutually subversive relation between representing the aspects of the string that figure in generalizations and representing its concatenation structure. In the end, neither is done satisfactorily.

Rumelhart and McClelland display some ambivalence about the Wickelfeature. At one point they dismiss the computational difficulty of recovering a string from a Wickelfeature set as one that is easily overcome by parallel processing "in biological hardware" (p. 262). At another point they show how the Wickelfeature-to-Wickelphone re-conversion can be done in a binding network that utilizes a certain genus of connectionist mechanisms, implying again that this process is to be taken seriously as part of the model. Yet they write (PDPII, p. 239):

> 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.

This disclaimer is at odds with the centrality of the Wickelfeature in the model's design. The Wickelfeature structure is not some kind of approximation that can easily be sharpened and refined; it is categorically the wrong kind of thing for the jobs assigned to it.[9] At the same time, the Wickelphone or something similar is demanded by the most radically distributed forms of distributed representations, which resolve order relations (like concatenation) into unordered sets of features. Without the Wickelphone, Rumelhart and McClelland have no account about how phonological strings are to be analyzed for significant patterning.

4.2. *Phonology and morphology*

The RM model maps from input to output in a single step, on the assumption that the past tense derives by a direct phonetic modification of the stem. The regular endings *-t, -d, -id*, make their appearance in the same way as the

[9] Compare in this regard certain other aspects of the model which are clearly inaccurate, but represent harmless oversimplifications. The actual set of phonetic features used to describe individual phones (p. 235) doesn't make enough distinctions for English, much less language at large, nor is it intended to; but the underlying strategy of featural analysis is solidly supported in the scientific literature. Similarly, the frequency classifications of the verbs in the study derive from the Kucera–Francis count over a written corpus, which shows obvious divergences from the input encountered by a learner (for examples, see footnote 24). Such aberrations, which have little impact on the model's behavior, could be corrected easily, with no structural re-design.

vowel changes *i → a (sing – sang)* or *u → o (choose – chose)*. Rumelhart and McClelland claim as an advantage of the model that "[a] uniform procedure is applied for producing the past-tense form in every case." (PDPII, p. 267) This sense of uniformity can be sustained, however, only if past tense formation is viewed in complete isolation from the rest of English phonology and morphology. We will show that Rumelhart and McClelland's very local uniformity must be paid for with extreme nonuniformity in the treatment of the broader patterns of the language.

The distribution of *t-d-id* follows a simple pattern: *id* goes after those stems ending in *t* or *d*; elsewhere, *t* (voiceless itself) goes after a voiceless segment and *d* (itself voiced) goes after a voiced segment. The real interest of this rule is that *none of it is specifically bound to the past tense*. The perfect/passive participle and the verbal adjective use the very same *t-d-id* scheme: *was kicked – was slugged – was patted; a kicked dog – a flogged horse – a patted cat*. These categories cannot be simply identified as copies of the past tense, because they have their own distinctive irregular formations. For example, past *drank* contrasts with the participle *drunk* and the verbal adjective *drunken*. Outside the verbal system entirely there is yet another process that uses the *t-d-id* suffix, with the variants distributed in exactly the same way as in the verb forms, to make adjectives from nouns, with the meaning 'having X' (Jespersen, 1942, p. 426 ff.):

(8)	-t	-d	-id
	hooked	long-nosed	one-handed
	saber-toothed	horned	talented
	pimple-faced	winged	kind-hearted
	foul-mouthed	moneyed	warm-blooded
	thick-necked	bad-tempered	bareheaded

The full generality of the component processes inherent in the *t-d-id* alternation only becomes apparent when we examine the widespread *s-z-iz* alternation found in the diverse morphological categories collected below:

(9)	Category	-s	-z	-iz
a.	Plural	hawks	dogs	hoses
b.	3psg	hits	sheds	chooses
c.	Possessive	Pat's	Fred's	George's
d.	has	Pat's	Fred's	George's
e.	is	Pat's	Fred's	George's
f.	does	what's	where's	–
g.	Affective	Pats(y)	Wills, bonkers	–
h.	adverbial	thereabouts	towards, nowadays	–
i.	Linking -s	huntsman	landsman	–

These 9 categories show syncretism in a big way—they use the same phonetic resources to express very different distinctions.

The regular noun plural exactly parallels the 3rd person singular marking of the verb, despite the fact that the two categories (noun/verb, singular/plural) have no notional overlap. The rule for choosing among *s-z-ɨz* is this: *ɨz* goes after stems ending in sibilants (*s,z,š,ž,č,ǰ*); elsewhere, *s* (itself voiceless) goes after voiceless segments, *z* (voiced itself) goes after voiced segments. The distribution of *s/z* is exactly the same as that of *t/d*. The rule for *ɨz* differs from that for *ɨd* only inasmuch as *z* differs from *d*. In both cases the rule functions to separate elements that are phonetically similar: as the sibilant *z* is to the sibilants, so the alveolar stop *d* is to the alveolar stops *t* and *d*.

The possessive marker and the fully reduced forms of the auxiliary *has* and the auxiliary/main verb *is* repeat the pattern. These three share the further interesting property that they attach not to nouns but to noun phrases, with the consequence that in ordinary colloquial speech they can end up on any kind of word at all, as shown in (10) below:

(10) a. [my mother-in-law]'s hat (cf. plural: *mothers-in-law*)
b. [the man you met]'s dog
c. [the man you spoke to]'s here. (Main verb *be*)
d. [the student who did well]'s being escorted home. (Auxiliary *be*)
e. [the patient who turned yellow]'s been getting better. (Auxiliary *has*)

The remaining formal categories (10f–h) share the *s/z* part of the pattern. The auxiliary *does*, when unstressed, can reduce colloquially to its final sibilant:[10]

[10] A post-sibilant environment in which *ɨz* would be necessary seems somewhat less available in natural speech:

(i) ? What church's he go to?
(ii) ?? Whose lunch's he eat from?
(iii) ?? Which's he like better?
(iv) ?? Whose's he actually prefer?

We suspect that the problem here lies in getting *does* to reduce at all in such structural environments, regardless of phonology. If this is right, then (i) and (ii) should be as good (or bad) as structurally identical (v) and (vi), where the sibilant-sibilant problem doesn't arise:

(v) ? What synagogue's he go to?
(vi) ? Whose dinner's he eat from?

Sentence forms (iii) and (iv) use the wh-determines *which* and *whose* without following head nouns, which may introduce sufficient additional structural complexity to inhibit reduction. At any rate, this detail, though interesting in itself, is orthogonal to the question of what happens to *does* when it does reduce.

(11) a. 'Z he like beans?
b. What's he eat for lunch?
c. Where's he go for dinner?

The affective marker *s/z* forms nicknames in some dialects and argots, as in *Wills* from *William*, *Pats* from *Patrick*, and also shows up in various emotionally-colored neologisms like *bonkers*, *bats*, paralleling *-y* or *-o* (*batty*, *wacko*), with which it sometimes combines (*Patsy*, *fatso*). A number of adverbial forms are marked by *s/z*—*unawares, nowadays, besides, backwards, here/there/whereabouts, amidships*. A final, quite sporadic (but phonologically regular) use links together elements of compounds, as in *huntsman, statesman, kinsman, bondsman*.

The reason that the voiced/voiceless choice is made identically throughout English morphology is not hard to find: it reflects the prevailing and inescapable phonetics of consonant cluster voicing in the language at large. Even in unanalyzable words, final obstruent clusters have a single value for the voicing feature; we find only words like these:

(12) a.	ax, fix, box	[ks]
b.	act, fact, product	[kt]
c.	traipse, lapse, corpse	[ps]
d.	apt, opt, abrupt	[pt]
e.	blitz, kibitz, Potts	[ts]
f.	post, ghost, list	[st]

Entirely absent are words ending in a cluster with mixed voicing: [*zt*], [*gs*], [*kz*], etc.[11] Notice that after vowels, liquids, and nasals (non-obstruents) a voicing contrast is permitted:

(13) a.	lens – fence	[nz] – [ns]
b.	furze – force	[rz] – [rs]
c.	wild – wilt	[ld] – [lt]
d.	bulb – help	[lb] – [lp]
e.	goad – goat	[od] – [ot]
f.	niece – sneeze	[is] – [iz]

If we are to achieve uniformity in the treatment of consonant-cluster voicing, we must not spread it out over 10 or so distinct morphological form generators (i.e., 10 different networks), and then repeat it once again in the phonetic component that applies to unanalyzable words. Otherwise, we

[11] In noncomplex words obstruent clusters are overwhelmingly voiceless: the word *adze* [*dz*] pretty much stands alone.

would have no explanation for why English contains, and why generation after generation of children easily learn, the exact same pattern eleven or so different times. Eleven unrelated sets of cluster patternings would be just as likely. Rather, the voicing pattern must be factored out of the morphology and allowed to stand on its own.

Let's see how the cross-categorial generalizations that govern the surface shape of English morphemes can be given their due in a rule system. Suppose the phonetic content of the past tense marker is just /d/ and that of the diverse morphemes in (9) is /z/. There is a set of morphological rules that say how morphemes are assembled into words: for example, Verb-past = stem + /d/; Noun-pl = stem + /z/; Verb-3psg = stem + /z/. Given this, we can invoke a single rule to derive the occurrences of [t] and [s]:

(14) *Voicing Assimilation*. Spread the value of voicing from one obstruent to the next in word final position.[12]

Rule (14) is motivated by the facts of simplex words shown above: it holds of *ax* and *adze* and is restricted so as to allow *goat* and *horse* to escape unaffected—they end in single obstruents, not clusters. When a final cluster comes about via morphology, the rule works like this:

(15)	a.	pig + /z/		Vacuous
	b.	pit + /z/	→	[pIts]
	c.	pea + /z/		No Change
	d.	rub + /d/		Vacuous
	e.	rip + /d/	→	[rIpt]
	f.	tow + /d/		No Change

The crucial effect of the rule is to devoice /d/ and /z/ after voiceless obstruents; after voiced obstruents its effect is vacuous and after nonobstruents—vowels, liquids, nasals—it doesn't apply at all, allowing the basic values to emerge unaltered.[13]

The environment of the variant with the reduced vowel *ɨ* is similarly constant across all morphological categories, entailing the same sort of uniform treatment. Here again the simplex forms in the English vocabulary provide the key to understanding: in no case are the phonetic sequences [tt], [dd], [sibilant-sibilant] tolerated at the end of unanalyzable words, or even inside

[12] More likely, syllable-final position.

[13] Notice that if /t/ and /s/ were taken as basic, we would require a special rule of voicing, restricted to suffixes, to handle the case of words ending in vowels, liquids, and nasals. For example, *pea* + /s/ would have to go to *pea* + [z], even though this pattern of voicing is not generally required in the language: cf. the morphologically simplex word *peace*. Positing /d/ and /z/ as basic, on the other hand, allows the rule (14), which is already part of English, to derive the suffixal voicing pattern without further ado.

them.[14] English has very strong general restrictions against the clustering of identical or highly similar consonants. These are not mere conventions deriving from vocabulary statistics, but real limitations on what native speakers of English have learned to pronounce. (Such sequences are allowed in other languages.) Consequently, forms like [skɪdd] from *skid* + /d/ or [ǰʌǰz] from *judge* + /z/ are quite impossible. To salvage them, a vowel comes in to separate the ending from a too-similar stem final consonant. We can informally state the rule as (16):

(16) *Vowel Insertion*. Word-finally, separate with the vowel *ɨ* adjacent consonants that are too similar in place and manner of articulation, as defined by the canons of English word phonology.

The two phonological rules have a competitive interaction. Words like *passes* [pæsɨz] and *pitted* [pɪtɨd] show that Vowel Insertion will always prevent Voicing Assimilation: from *pass* + /z/ and *pit* + /d/ we never get [pæsɨs] or [pɪtɨt], with assimilation to the voiceless final consonant. Various lines of explanation might be pursued; we tentatively suggest that the outcome of the competition follows from the rather different character of the two rules. Voicing Assimilation is highly phonetic in character, and might well be part of the system that implements phonological representations rather than part of the phonology proper, where representations are defined, constructed, and changed. If Vowel Insertion, as seems likely, actually changes the representation prior to implementation, then it is truly phonological in character. Assuming the componential organization of the whole system portrayed above, with a flow between components in the direction Morphology → Phonology → Phonetics, the pieces of the system fall naturally into place. Morphology provides the basic structure of stem + suffix. Phonology makes various representational adjustments, including Vowel Insertion, and Phonetics then implements the representations. In this scheme, Voicing Assimilation, sitting in the phonetic component, never sees the suffix as adjacent to a too-similar stem-final consonant.

Whatever the ultimate fate of the details of the competition, it is abundantly clear that the English system turns on a fundamental distinction between phonology and morphology. Essential phonological and phonetic processes are entirely insensitive to the specifics of morphological composition and sweep across categories with no regard for their semantic or syntactic content. Such processes define equivalences at one level over items that are distinct at the level of phonetics: for English suffixes, *t* = *d* = *ɨd* and *s* = *z*

[14] This is of course a phonological restriction, not an orthographic one. The words *petty* and *pity*, for example, have identical consonantal phonology.

= *iz*. As a consequence, the learner infers that there is one suffix for the regular past, not three; and one suffix, not three, for each of plural, 3rd person singular, possessive, and so on. The phonetic differences emerge automatically; as would be expected in such cases, uninstructed native speakers typically have no awareness of them.

Rumelhart and McClelland's pattern associator is hobbled by a doctrine we might dub "morphological localism": the assumption that there is for each morphological category an encapsulated system that handles every detail of its phonetics. This they mischaracterize as a theoretically desirable "uniformity". In fact, morphological localism destroys uniformity by preventing generalization across categories and by excluding inference based on larger-scale regularities. Thus it is inconsistent with the fact that the languages that people learn are shaped by these generalizations and inferences.

The shape of the system. It is instructive to note that although the various English morphemes discussed earlier all participate in the general phonological patterns of the language, like the past tense they can also display their own particularities and subpatterns. The 3rd person singular is extremely regular, with a few lexical irregularities (*is, has, does, says*) and a lexical class (modal auxiliaries) that can't be inflected (*can, will*, etc.). The plural has a minuscule number of non-/z/ forms (*oxen, children, geese, mice*, ...), a Ø suffixing class (*sheep, deer*), and a fricative-voicing subclass (*leaf-leaves, wreath-wreathes*). The possessive admits no lexical peculiarities (outside of the pronouns), presumably because it adds to phrases rather than lexical items, but it is lost after plural /z/ (*men's* vs. *dogs'*) and sporadically after other *z*'s. The fully reduced forms of *is* and *has* admit no lexical or morphologically-based peculiarities whatever, presumably because they are syntactic rather than lexical.

From these observations, we can put together a general picture of how the morphological system works. There are some embracing regularities:

1. All inflectional morphology is suffixing.
2. All nonsyllabic regular suffixes are formed from the phonetic substance /d/ or /z/; that is, they must be the same up to the one feature distinguishing *d* from *z*: sibilance.
3. All morphemes are liable to re-shaping by phonology and phonetics.
4. Categories, inasmuch as they are lexical, can support specific lexical peculiarities and subpatterns; inasmuch as they are nonlexical, they must be entirely regular.

Properties (1) and (2) are clearly English-bound generalizations, to be learned by the native speaker. Properties (3) and (4) are replicated from language to language and should therefore be referred to the general capacities of the learner rather than to the accidents of English. Notice that we have lived up to our promise to show that the rules governing the regular past tense are not idiosyncratic to it: beyond even the phonology discussed above, its intrinsic phonetic content is shared up to one feature with the other regular nonsyllabic suffixes; and the rule of inflectional suffixation itself is shared generally across categories. We have found a highly modular system, in which the mapping from uninflected stem to the phonetic representation of the past tense form breaks down into a cascade of independent rule systems, and each rule system treats its inputs identically regardless of how they were originally created.

It is a nontrivial problem to design a device that arrives at this characterization on its own. An unanalyzed single module like the RM pattern associator that maps from features to features cannot do so.

4.3. Lexical items

The notion of a 'word' or 'morpheme' is so basic to our intuitive understanding of language that it is easy to forget the role it plays in systematic linguistic explanation. As a result, use of the representational structure known as a 'lexical item' might be seen as mere tradition, and one of the revolutionary aspects of the RM model—that it contains nothing corresponding to a lexical item other than its phonetic composition—might be dismissed as a harmless iconoclasm. Here we show that, on the contrary, lexical items as explicit representations play a crucial role in many linguistic phenomena.

4.3.1. Preservation of stem and affix

The pattern associator suffers from a fundamental design problem which prevents it from truly grasping even the simplest morphological generalization. Because the relation between stem and past tense is portrayed as a transduction from one low-level featural representation to another, literally replacing every feature in the input, it becomes an inexplicable accident that the regular formation rule preserves the stem unaltered. The identity map has no cachet in the pattern associator; it is one among very many (including the reverse map) that happen to produce strings in the output. Yet a tendency toward preservation of stem identity, a typical linguistic phenomenon, is an immediate consequence of the existence of morphology as a level of description: if the rule is Word = Stem + Affix, then ceteris paribus the stem comes through. What makes ceteris not exactly paribus is the potential existence of phonological and phonetic accommodations, but even these will be relatively minute in a properly formulated theory.

The other side of the morphological coin is the preservation of *affix* identity. The suffixal variants *t* and *d* are matched with *id*, not with *iz* or *oz* or *og* or any other conceivable but phonetically distant form. Similarly, morphemes which show the *s/z* variants take *iz* in the appropriate circumstances, not *id* or *od* or *gu*. This follows directly from our hypothesis that the morphemes in question have just one basic phonetic content—/d/ or /z/—which is subject to minor contextual adjustments. The RM model, however, cannot grasp this generalization. To see this, consider the Wickelphone map involved in the *id* case, using the verb *melt* as an example:

(17) a. {#me, mel, elt, lt#} → {#me, mel, elt, lti, tid, id#}
b. lt# → lti, tid, id#

The replacement Wickelphone (or more properly—Wickelfeature set) *id#* has no relation to the stem-final consonant and could just as well be *iz#* or *ig#*. Thus the RM model cannot explain the prevalence across languages of inflectional alternations that preserve stem and affix identities.

4.3.2. Operations on lexical items

The generalizations that the RM model extracts consist of specific correlations between particular phone sequences in the stem and particular phone sequences in the past form. Since the model contains no symbol corresponding to a stem per se, independent of the particular phone sequences that happen to have exemplified the majority of stems in the model's history, it cannot make any generalization that refers to stems per se, cutting across their individual phonetic contents. Thus a morphological process like reduplication, which in many languages copies an entire stem (e.g. yielding forms roughly analogous to *dum-dum* and *boom-boom*), cannot be acquired in its fully general form by the network. In many cases it can "memorize" *particular* patterns of reduplication, consisting of mappings between particular feature sequences and their reduplicated counterparts (though even here problems can arise because of the poverty of the Wickelfeature representation, as we pointed out in discussing Wickelphonology), but the concept "Copy the stem" itself is unlearnable; there is no unitary representation of a thing to be copied and no operation consisting of copying a variable regardless of its specific content. Thus when a new stem comes in that does not share many features with the ones encountered previously, it will not match any stored patterns and reduplication will not apply to it.[15]

[15] It is worth noting that reduplication, which always calls on a variable (if not 'stem', then 'syllable' or 'foot') is one of the most commonly used strategies of word-formation. In one form or another, it's found in hundreds, probably thousands, of the world's languages. For detailed analysis, see McCarthy and Prince (forthcoming).

The point strikes closer to home as well. The English regular past tense rule adds an affix to a stem. The rule doesn't care about the contents of the stem; it mentions a variable, "stem", that is cashed in independently for information stored in particular lexical entries. Thus the rule, once learned, can apply across the board independent of the set of stems encountered in the learner's history. The RM model, on the other hand, learns the past tense alternation by linking phonetic features of inflected forms directly to the particular affix features of the stem (for example, in *pat* – *patted* the *id#* Wickelfeatures are linked directly to the entire set of features for *pat*: *#pæ*, *pæt*, etc.). Though much of the activation for the affix features eventually is contributed by some stem features that cut across many individual stems, such as those at the end of a word, not all of it is; some contribution from the word-specific stem features that are well-represented in the input sample can play a role as well. Thus the RM model could fail to generate any past tense form for a new stem if the stem did not share enough features with those stems that were encountered in the past and that thus grew their own strong links with past tense features. When we examine the performance of the RM model, we will see how some of its failures can probably be attributed to the fact that what it learns is associated with particular phone sequences as opposed to variables standing for stems in general.

4.3.3. Lexical items as the locus of idiosyncrasy

For the RM model, membership in the strong classes is determined entirely by phonological criteria; there is no notion of a "lexical item", as distinct from the phone-sequences that make up the item, to which an 'irregular' tag can be affixed. In assessing their model, Rumelhart and McClelland write:

> 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. (PDPII, p. 267)

If Rumelhart and McClelland are right, there can be no homophony between regular and irregular verbs or between items in distinct irregular classes, because words are nothing but phone-sequences, and irregular forms are tied directly to these sequences. This basic empirical claim is transparently false. Within the strong class itself, there is a contrast between *ring* (past: *rang*) and *wring* (past: *wrung*) which are only orthographically distinct. Looking at the broader population, we find the string *lay* shared by the items *lie* (past: *lied*) 'prevaricate' and *lie* (past: *lay*) 'assume a recumbent position'. In many dialects, regular *hang* refers to a form of execution, strong *hang* means merely 'suspend'. One verb *fit* is regular, meaning 'adjust'; the other, which refers to the shape-or-size appropriateness of its subject, can be strong:

(18) a. That shirt never fit/?fitted me.
b. The tailor fitted/*fit me with a shirt.

The sequence [kʌm] belongs to the strong system when it spells the morpheme *come*, not otherwise: contrast *become, overcome* with *succumb, encumber.*

An excellent source of counterexamples to the claim that past tense formation collapses the distinctions between words and their featural decomposition is supplied by verbs derived from other categories (like nouns or adjectives). The significance of these examples, which were first noticed in Mencken (1936), has been explored in Kiparsky (1982a, b)[16]

(19) a. He braked the car suddenly. ≠ broke
b. He flied out to center field. ≠ flew
c. He ringed the city with artillery. *rang
d. Martina 2-setted Chris. *2-set
e. He subletted/sublet the apartment.
f. He sleighed down the hill. *slew
g. He de-flea'd his dog. *de-fled
h. He spitted the pig. *spat
i. He righted the boat. *rote
j. He high-sticked the goalie. *high-stuck
k. He grandstanded to the crowd. *grandstood.

This phenomenon becomes intelligible if we assume that irregularity is a property of verb *roots*. Nouns and adjectives by their very nature do not classify as irregular (or regular) with respect to the past tense, a purely verbal notion. Making a noun into a verb, which is done quite freely in English, cannot produce a new verb root, just a new verb. Such verbs can receive no special treatment and are inflected in accord with the regular system, regardless of any phonetic resemblance to strong roots.

In some cases, there is a circuitous path of derivation: V → N → V. But the end product, having passed through nounhood, must be regular no matter what the status of the original source verb. (By "derivation" we refer to relations intuitively grasped by the native speaker, not to historical etymology.) The baseball verb *to fly out*, meaning 'make an out by hitting a fly ball that gets caught', is derived from the baseball noun *fly (ball)*, meaning 'ball hit on a conspicuously parabolic trajectory', which is in turn related to the simple strong verb *fly* 'proceed through the air'. Everyone says "he flied out"; no mere mortal has yet been observed to have "flown out" to left field.

[16] Examples (19b) and (h) are from Kiparsky.

Similarly, the noun *stand* in the lexical compound *grandstand* is surely felt by speakers to be related to the homophonous strong verb, but once made a noun its verbal irregularity cannot be resurrected: **he grandstood*. A derived noun cannot retain any verbal properties of its base, like irregular tense formation, because nouns in general can't have properties such as tense. Thus it is not simply derivation that erases idiosyncrasy, but departure from the verb class: *stand* retains its verbal integrity in the verbs *withstand*, *understand*, as *throw* does in the verbs *overthrow*, *underthrow*.[17] Kiparsky (1982a, b) has pointed out that regularization-by-derivation is quite general and shows up wherever irregularity is to be found. In nouns, for example, we have the *Toronto Maple Leafs*, not **Leaves*, because use in a name strips a morpheme of its original content. Similar patterns of regularization are observed very widely in the world's languages.

One might be tempted to try to explain these phenomena in terms of the meanings of regular and irregular versions of a verb. For example, Lakoff (1987) appeals to the distinction between the 'central' and 'extended' senses of polysemous words, and claims that irregularity attaches only to the 'central sense' of an item. It is a remarkable fact—indeed, an insult to any naive idea that linguistic form is driven by meaning—that polysemy is irrelevant to the regularization phenomenon. Lakoff's proposed generalization is not sound. Consider these examples:

(20) a. He wetted his pants. *wet* regular in central sense.
b. He wet his pants. *wet* irregular in extended sense.

(21) a. They heaved the bottle overboard. *heave* regular in central sense.
b. They hove to. *heave* irregular in extended sense.

It appears that a low-frequency irregular can occasionally become locked into a highly specific use, regardless of whether the sense involved is 'central' or 'extended'. Thus the purely semantic or metaphorical aspect of sense extension has no predictive power whatsoever. Verbs like *come, go, do, have, set, get, put, stand* ... are magnificently polysemous (and become more so in combination with particles like *in, out, up, off*), yet they march in lockstep

[17] When the verb is the 'head' of the word it belongs to, it passes on its categorial features to the whole word, including both verb-ness and more specialized morphological properties like irregularity (Williams, 1981). Deverbal nouns $[_{N}V]$ and denominal verbs $[_{V}N]$ must therefore be headless, whereas prefixed verbs are headed $[_{V}\text{PREF-V}]$. Notice that there can be uncertainty and dialect differences in the interpretation of individual cases. The verb *sublet* can be thought of as denominal, $[_{V}[_{N}\text{sublet}]]$ giving *subletted*, or as a prefixed form headed by the verb *to let*, giving past tense *sublet*.

through the same nonregular paradigms in central and extended senses—regardless of how strained or opaque the metaphor.[18] Similarly, they retain their nonregular forms when combined with bound affixes that recur in word-formation patterns in the language, even if the meaning of the whole is not composed of the meaning of its parts: *forget/forgot, forgive/forgave, understand/understood, undertake/undertook, overcome/overcame* (see Aronoff, 1976, for other examples of this kind of phenomenon). But when a verb is transparently derived from a noun or adjective, the irregular system is *predictably* by-passed. The critical factors are lexical category in the formal sense—noun, verb, adjective—and the structural analysis of the word into entities such as root, stem, head, prefix, which are purely and autonomously morphological.

To master the actual system, then, the learner must have access to lexical information about each item, ranging from its derivational status (is the item a primitive root? is it derived from a noun or another verb?) to its specific lexical identity (is the item at hand *ring* or *wring*, $hang_1$ or $hang_2$, lie_1 or lie_2, etc.?). The RM model does without the notion 'lexical item' at the cost of major lapses in accuracy and coverage.

Our basic finding is independent of how the notion 'lexical item' is implemented. If a lexical item is a distributed pattern of activation—that is to say, just a set of semantic, syntactic, morphological, and phonological features—it remains true that past tense formation must be sensitive to various aspects of the pattern. It is hardly acceptable, however, to allow past tense formation (or morphology in general) to access every scrap of lexical information. Categorial information like root vs. derived status figures in the morphology of language after language, and with comparable effects, whereas the specific semantic distinctions between, say, *ring* and *wring* are hardly the basis for any real generalization. (Such verbs could have their class assignments reversed with no consequences for the rest of the language. We return to this point in Sections 8.3.1 and 8.3.4.) What's important is that *ring* ≠ *wring*, $hang_1 \neq hang_2$; that they are not the same items. From such cases, it is clear that classification is not driven by any particular feature of the lexical item; rather, arbitrary assignment to a strong class is *itself* a lexical feature. Because morphology is sensitive to gross distinctness (α is not the same as β)

[18] Compare in this regard Ross's (1975) study of productive affixation, which uncovers an actual constraint involving the central/extended distinction. Ross finds that prefixes like *re-*, *un-*, *mis-*, which affect meaning, are sensitive in various ways to the meaning of the base they attach to. He amply documents the fact that such prefixes reject metaphorically extended bases. Thus: "Horace Silver (*re-)cut Liberace" (*cut* = 'played better than'), "Larry (*mis-)fed Dennis" (*fed* = 'passed the basketball to'). Examination of Ross's numerous examples shows *not one* where metaphorical extension affects irregularity. The contrast could not be starker. Notions like 'past tense form' have no systematic sensitivity to the lexical semantics of the base.

rather than to every possible semantic, syntactic, and pragmatic fillip, we can conclude that lexical items do indeed possess an accessible local identity as well as a distributed featural decomposition.

4.4. *The strong system and the regular system*

The RM model embodies the claim that the distinction between regular and irregular modes of formation is spurious. At this point, we have established the incorrectness of two assumptions that were supposed to support the broader claim of uniformity.

> Assumption #1. "All past tenses are formed by direct phonetic modification of the stem." We have shown that the regular forms are derived through affixation followed by phonological and phonetic adjustment.
>
> Assumption #2. "The inflectional class of any verb (regular, subregular, irregular) can be determined from its phonological representation alone." We have seen that membership in the strong classes depends on lexical and morphological information.

These results still leave open the question of disparity between the regular and strong systems. To resolve it, we need a firmer understanding of how the strong system works. We will find that the strong system has a number of distinctive peculiarities which are related to its being a partly structured list of exceptions. We will examine five:

1. Phonetic similarity criteria on class membership.
2. Prototypicality structure of classes.
3. Lexical distinctness of stem and past tense forms.
4. Failure of predictability in verb categorization.
5. Lack of phonological motivation for the strong-class changes.

4.4.1. *Hypersimilarity*

The strong classes are often held together, if not exactly defined, by phonetic similarity. The most pervasive constraint is monosyllabism: 90% of the strong verbs are monosyllabic, and the rest are composed of a monosyllable combined with an unstressed and essentially meaningless prefix.[19]

[19] The polysyllabic strong verbs are:
arise, awake
become, befall, beget, begin, behold, beset, beshit, bespeak
forbear, forbid, forget, forgive, forgo, forsake, forswear, foretell
mistake
partake

(*continued*) →

Within the various classes, there are often significant additional resemblances holding between the members. Consider the following sample of typical classes, arranged by pattern of change in past and past participle ("*x–y–z*" will mean that the verb has the vowel *x* in its stem, *y* in its past tense form, and *z* in its past participle). Our judgments about the cited forms are indicated as follows: *?Verb* means that usage of the irregular past form of *Verb* is somewhat less natural than usual; *??Verb* means that *Verb* is archaic or recherché-sounding in the past tense.

(22) Some strong verb types
a. x - [u] - x(o)+n
blow, grow, know, throw
draw, withdraw
fly
??slay
b. [e] - [U] - [e]+en
take, mistake, forsake, shake
c. [ay] - [aw] - [aw]
bind, find, grind, wind
d. [d] - [t] - [t]
bend, send, spend, ?lend, ??rend
build
e. [ɛ] - [ɔ] - [ɔ]+n
swear, tear, wear, ?bear, ??forswear, ??forbear
get, forget, ??beget
?tread

The members of these classes share much more than just a pattern of changes. In the *blow*-group (22a), for example, the stem-vowel becomes [u] in the past; this change could in principle apply to all sorts of stems, but in fact the participating stems are all vowel-final, and all but *know* begin with a CC cluster. In the *find*-group (22c) the vowel change [ay] → [aw] could apply to any stem in [ay], but it only applies to a few ending in [nd]. The change of [d] to [t] in (22d) occurs only after sonorants [n, l] and mostly when the stem rhymes in *-end*. Rhyming is also important in (22b), where every-

understand, undergo
upset
withdraw, withstand
The prefixes *a-*, *be-*, *for-*, *under-*, *with-* do not carry any particular meaning, nor in fact do most of the stems. (There is nothing about 'for' and 'get', for example, that helps us interpret *forget*.) Their independent existence in other forms is sufficient to support a sense of compositeness; see Aronoff (1976). As mentioned, this shows that morphology is in some sense a separate, abstract component of language.

thing ends in *-ake* (and the base also begins with a coronal consonant), and in (22e), where *-ear* has a run.

Most of the multi-verb classes in the system are in fact organized around clusters of words that rhyme and share other structural similarities, which we will call *hypersimilarities*. (The interested reader is referred to the Appendix for a complete listing.) The regular system shows no signs of such organization. As we have seen, the regular morpheme can add onto any phonetic form—even those most heavily tied to the strong system, as long as the lexical item involved is not a primary verb root.

4.4.2. Prototypicality

The strong classes often have a kind of prototypicality structure. Along the phonetic dimension, Bybee and Slobin (1982) point out that class cohesion can involve somewhat disjunctive 'family resemblances' rather than satisfaction of a strict set of criteria. In the *blow*-class (22a), for example, the central exemplars are *blow, grow, throw*, all of the form [CRo], where R is a sonorant. The verb *know* [no] lacks the initial C in the modern language, but otherwise behaves like the exemplars. The stems *draw* [drɔ] and *slay* [sle] fit a slightly generalized pattern [CRV] and take the generalized pattern x–u–x in place of o–u–o. The verb *fly* [flay] has the diphthong [ay] for the vowel slot in [CRV], which is unsurprising in the context of English phonology, but unlike *draw* and *slay* it takes the concrete pattern of changes in the exemplars: x–u–o rather than x–u–x. Finally, all take *-n* in the past participle.

Another kind of prototypicality has to do with the degree to which strong forms allow regular variants. (This need not correlate with phonetic centrality—notice that all the words in the *blow*-class are quite secure in their irregular status.) Consider the class of verbs which add -t and lax the stem-vowel:

(23) V: - V - V(+t)
keep, sleep, sweep, weep (?weeped/wept), creep (?creeped/crept), leap (leaped/leapt)
feel, deal (?dealed/dealt), kneel (kneeled/?knelt)
mean
dream (dreamed/?dreamt)
leave
lose

Notice the hypersimilarities uniting the class: the almost exclusive prevalence of the vowel [i]; the importance of the terminations [-ip] and [-il].

The parenthetical material contains coexisting variants of the past forms that, according to our judgments, are acceptable to varying degrees. The range of prototypicality runs from 'can only be strong' (*keep*) through 'may be either' (*leap*) to 'may possibly be strong' (*dream*). The source of such variability is probably the low but nonzero frequency of the irregular form, often due to the existence of conflicting but equally high-status dialects (see Bybee, 1985).

The regular system, on the other hand, does not have prototypical exemplars and does not have a gradient of variation of category membership defined by dimensions of similarity. For example, there appears to be no sense in which *walked* is a better or worse example of the past tense form of *walk* than *genuflected* is of *genuflect*. In the case at hand, there is no reason to assume that regular verbs such as *peep, reap* function as a particularly powerful attracting cluster, pulling *weep, creep, leap* away from irregularity. Historically, we can clearly see attraction in the opposite direction: according to the *OED*, *knelt* appears first in the 19th century; such regular verbs as *heal, peel, peal, reel, seal, squeal* failed to protect it; as regular forms they could not do so, on our account, because their phonetic similarity is not perceived as relevant to their choice of inflection, so they do not form an attracting cluster.

4.4.3. Lexicality

The behavior of low-frequency forms suggests that the stem and its strong past are actually regarded as distinct lexical items, while a regular stem and its inflected forms, no matter how rare, are regarded as expressions of a single item.

Consider the verb *forgo*: though uncommon, it retains a certain liveliness, particularly in the sarcastic phrase "forgo the pleasure of ...". The past tense must surely be *forwent* rather than **forgoed*, but it seems entirely unusable. Contrast the following example, due to Jane Grimshaw:

(24) a. *Last night I forwent the pleasure of grading student papers.
b. You will excuse me if I forgo the pleasure of reading your paper until it's published.

Similarly but more subtly, we find a difference in naturalness between stem and past tense when the verbs *bear* and *stand* mean 'tolerate':

(25) a. I don't know how she bears it.
b. (?) I don't know how she bore it.
c. I don't know how she stands him.
d. (?) I don't know how she stood him.

The verb *rend* enjoys a marginal subsistence in the phrase *rend the fabric of society*, yet the past seems slightly odd: *The Vietnam War rent the fabric of American society*. The implication is that familiarity can accrue differentially to stem and past tense forms; the use of one in a given context does not always entail the naturalness of the other.

This phenomenon appears to be absent from the regular system. There are regular verbs that are trapped in a narrow range of idioms, like *eke* in "eke out", *crook* in "crook one's finger", *stint* in "stint no effort", yet all inflected forms seem equivalent. Furthermore, rare or self-conscious verbs like *anastomose, fleech, fleer, incommode, prescind* show no further increment of oddness or uncertainty in the past tense. Suppose that it is only the items actually listed in the lexicon that gain familiarity, rather than each individual inflected form. If regular forms are rule-generated from a single listed item, then all forms should freely inherit statistics from each other. Irregular forms, on the other hand, listed because unpredictable, should be able to part company even if they belong to a single paradigm.

4.4.4. Failures of predictability

Even when a verb matches the characteristic patterns of any of the classes in the strong system, no matter how closely, there can be no guarantee that the verb will be strong. If the verb is strong, its similarity to the characteristic patterns of the subclasses cannot always predict which of these subclasses it will fall into. Verbs like *flow, glow, crow* are as similar to the words in the set *blow, grow, throw, know* as the members of the set are to each other; yet they remain regular. (Indeed, *crow* has turned regular in the last few hundred years.) As for subcategorization into one of the strong subclasses, consider the clear subregularity associated with the [I - æ - ʌ] and [I - ʌ - ʌ] vowel-change classes:

(26) a. I - æ - ʌ
ring, sing, spring
drink, shrink, sink, stink
swim
begin, spin, win
⟨run⟩

b. I - ʌ - ʌ
cling, sling, sting, string, swing, wring, fling (?flinged/flung),
slink (slinked/?slunk)
stick
dig
⟨hang⟩

The core members of these related classes end in *-ing* and *-ink*. (Bybee and Slobin note the family resemblance structure here, whereby the hallmark 'velar nasal' accommodates mere nasals on the one side (*swim*, etc.) and mere velars on the other (*stick, dig*); the stems *run* and *hang* differ from the norm in a vowel feature or two, as well.) Interestingly, no primitive English monosyllabic verb root that ends in *-ing* is regular. Forms like *ding, ping, zing*, which show no attraction to class (26), are tainted by onomatopoetic origins; forms like *ring* (surround), *king* (as in checkers), and *wing* are obviously derived from nouns. Thus the *-ing* class of verbs is the closest we have in English to a class that can be uniformly and, possibly, productively, inflected with anything other than the regular ending. Nevertheless, even for this subclass it is impossible to predict the actual forms from the fact of irregularity: *ring–rang* contrasts with *wring–wrung*; *spring–sprang* with *string–strung*; and *bring* belongs to an entirely unrelated class. This observation indicates that learners can pick up the general distinction regular/irregular at some remove from the particular patterns.

The regular system, in contrast, offers complete predictability.

4.4.5. Lack of phonological motivation for morphological rules

The rules that determine the shape of the regular morphemes of English are examples of true phonological (or even phonetic) rules: they examine a narrow window of the string and make a small-scale change. Such rules have necessary and sufficient conditions, which must be satisfied by elements present in the window under examination in order for the rule to apply. The conditioning factors are intrinsically connected with the change performed. Voicelessness in the English suffixes directly reflects the voicelessness of the stem-final consonant. Insertion of the vowel ɨ resolves the inadmissible adjacency of (what English speakers regard as) excessively similar consonants.

The relations between stem and past tense in the various strong verb classes are defined on phonological substance, but the factors affecting the relationship are not like those found in true phonological rules. In particular, the changes are for the most part entirely unmotivated by phonological conditions in the string. There is nothing in the environment *b_nd* that encourages [ay] to become [aw]; nothing about [CRo], the basic scheme of the *blow-* class, that causes a change to [CRu] or makes such a change more likely than in some other environment. These are arbitrary though easily definable changes tied arbitrarily to certain canonical forms, in order to mark an abstract morphological category: past tense. The patterns of similarity binding the classes together actually play no causal role in determining the changes that occur. A powerful association may exist, but it is merely conventional and could quite easily be otherwise (and indeed in the different dialects of the language spoken now or in the past, there are many different systems). Similarity relations serve essentially to qualify entry into a strong class rather than to provide an environment that causes a rule to happen.

There is one region of the strong system where discernibly phonological factors do play a role: the treatment of stems ending in [-t] and [-d]. No strong verb takes the suffix *id* (*bled/*bledded, got/*gotted*); the illicit cluster that would be created by suffixing /d/ is resolved instead by eliminating the suffix. This is a strategy that closely resembles the phonological process of degemination (simplification of identical adjacent consonants to a single consonant), which is active elsewhere in English. Nevertheless, if we examine the class of affected items, we see the same arbitrariness, prototypicality, and incomplete predictiveness we have found above. Consider the "no-change" class, which uses a single form for stem, past tense, and past participle—by far the largest single class of strong verbs, with about 25 members. In these examples, a word preceded by '?' has no natural-sounding past tense form in our dialect; words followed by two alternatives in parentheses have two possible forms, often with one of them (indicated by '?') worse-sounding than the other:

(27) No-change verbs
hit, slit, split, quit, spit (spit/spat), knit (knitted/?knit), ?shit, ??beshit
bid, rid
shed, spread, wed
let, set, upset, ?beset, wet (wetted/wet)
cut, shut
put
burst, cast, cost
thrust (thrusted/thrust), hurt

Although ending in [-t, d] is a necessary condition for no-change status, it is by no means sufficient. First of all, the general constraint of monosyllabism applies, even though it is irrelevant to degemination. Second, there is a strong favoritism for the vowels [I] and [ɛ], followed by a single consonant; again, this is of no conceivable relevance to a truly phonological process simplifying [td] and [dd] to [t] and [d]. Absent from the class, and under no attraction to it, are such verbs as *bat, chat, pat, scat*, as well as *jot, rot, spot, trot*, with the wrong sort of vocalism; and *dart, fart, smart, start, thwart, snort, sort, halt, pant, rant, want* with nonprototypical vowel and consonant structure. Even in the core class, we find arbitrary exceptions: *flit, twit, knit* are all regular, as are *fret, sweat, whet*, and some uses of *wet*. Beside strong *cut* and *shut*, we find regular *butt, jut, strut*. Beside *hurt* we find *blurt, spurt*; beside *burst*, we find regular *bust*. The phonological constraints on the class far exceed anything relevant to degemination, but in the end they characterize rather than define the class, just as we have come to expect.

Morphological classification responds to fairly large-scale measures on word structure: is the word a monosyllable? does it rhyme with a key exemplar? does it alliterate (begin with a similar consonant cluster) as an exemplar? Phonological rules look for different and much more local configurations: is this segment an obstruent that follows a voiceless consonant? are these adjacent consonants nearly identical in articulation? In many ways, the two vocabularies are kept distinct: we are not likely to find a morphological subclass holding together because its members each contain somewhere inside them a pair of adjacent obstruents; nor will we find a rule of voicing-spread that applies only in rhyming monosyllables. If an analytical engine is to generalize effectively over language data, it can ill afford to look upon morphological classification and phonological rules as processes of the same formal type.

4.4.6. Default structure

We have found major differences between the strong system and the regular system, supporting the view that the strong system is a cluster of irregular patterns, with only the *-ing* forms and perhaps the no-change forms displaying some active life as partially generalizable subregularities in the adult language. Membership in the strong system is governed by several criteria: (1) monosyllabism; (2) nonderived verb root status; (3) for the subregularities, resemblance to key exemplars. This means that the system is largely closed, particularly because verb roots very rarely enter the language (new verbs are common enough, but are usually derived from nouns, adjectives, or onomatopoetic expressions). At a few points in history, there have been borrowed items that have met all the criteria: *quit* and *cost* are both from French, for example (Jespersen, 1942). The regular system is free from such constraint. No canonical structure is required—for example, 'not a monosyllable'. No information about derivational status is required, such as 'must not be derived from an adjective'. Phonetic similarity to an exemplar plays no role either. Furthermore, the behavior of regular verbs is entirely predictable on general grounds. The regular rule of formation is an extremely simple default with very few characteristics of its own—perhaps only one, as we suggest above: that the morpheme is a stop rather than a fricative.

The regular system also has an internal default structure that is worthy of note, since it contrasts with the RM model's propensities. The rule Past = stem + /d/ covers all possible cases. Under narrowly defined circumstances, some phonology takes place: a vowel intrudes to separate stem and affix, voicelessness propagates from the stem. Elsewhere—the default case—nothing happens. It appears that language learners are fond of such architectures, which appear repeatedly in languages. (Indeed, in the history of English all inflection heads in this direction.) Yet the RM network, unlike the rule theory, offers us no insight. The network is equally able to learn a set of scattered, nonlocal, phonetically unrelated subregularities: for example, "suffix *t* if the word begins with *b*; prefix *ik* if the word ends in a vowel; change all *s*'s to *r* before *ta*"; etc. The RM model treats the regular class as a kind of fortuitously overpopulated subregularity; indeed, as three such classes, since the *d-t-id* alternation is treated on a par with the choice between strong subclasses. The extreme and categorical uniformity of the regular system disappears from sight, and with it the hope of identifying such uniformity as a benchmark of linguistic generalization.

4.4.7. *Why are the regular and strong systems so different?*

We have argued that the regular and strong systems have very different properties: the regular system obeys a categorical rule that is stated in a form that can apply to any word and that is adjusted only by very general phonological regularities; whereas the strong system consists of a set of subclasses held together by phonologically-unpredictable hypersimilarities which are neither necessary nor sufficient criteria for membership in the classes.

Why are they so different? We think the answer comes from the common-sense characterization of the psychological difference between regular and strong verbs. The past tense forms of strong verbs must be memorized; the past tense forms of regular verbs can be generated by rule. Thus the irregular forms are roughly where grammar leaves off and memory begins. Whatever affects human memory in general will shape the properties of the strong class, but not the regular class, by a kind of Darwinian selection process, because only the easily-memorized strong forms will survive. The 10 most frequent verbs of English are strong, and it has long been noted that as the frequency of a strong form declines historically, the verb becomes more likely to regularize. The standard explanation is that you can only learn a strong past by hearing it and only if you hear it often enough are you likely to remember it. However, it is important to note that the bulk of the strong verbs are of no more than middling frequency and some of them are actually rare, raising the question of how they managed to endure. The hypersimilarities and graded membership structure of the strong class might provide an answer. Rosch and Mervis (1975) note that conceptual categories, such as vegetables or tools, tend to consist of members with family resemblances to one another along a set of dimensions and graded membership determined by similarity to a prototype. They also showed in two experiments that it is easier for subjects to memorize the members of an artificial category if those members display a family resemblance structure than if they are grouped into categories arbitrarily. Since strong verbs, like Rosch and Mervis's artificial exemplars, must be learned one by one, it is reasonable to expect that the ones that survive, particularly in the middle and low frequencies, will be those displaying a family resemblance structure. In order words, the reason that strong verbs are either frequent or members of families is that strong verbs are memorized and frequency and family resemblance assist memorization.

The regular system must answer to an entirely different set of requirements: the rule must allow the user to compute the past tense form of any regular verb and so must be generally applicable, predictable in its output, and so on.

While it is possible that connectionist models of category formation (e.g. McClelland & Rumelhart, 1985) might offer insights into why family resemblance fosters category formation, it is the difference between fuzzy families of memorized exemplars and formal rules that the models leave unexplained.[20] Rumelhart and McClelland's failure to distinguish between mnemonics and productive morphology leads to the lowest-common-denominator 'uniformity' of accomplishing all change through arbitrary Wickelfeature replacement, and thus vitiates the use of psychological principles to explain linguistic regularities.

5. How good is the model's performance?

The bottom-line and most easily grasped claim of the RM model is that it succeeds at its assigned task: producing the correct past tense form. Rumelhart and McClelland are admirably open with their test data, so we can evaluate the model's achievement quite directly.

Rumelhart and McClelland submitted 72 new regular verbs to the trained model and submitted each of the resulting activated Wickelfeature vectors to the unconstrained whole-string binding network to obtain the analog of freely-generated responses. The model does not really 'decide' on a unique past tense form and stick with it thereafter; several candidates get strength values assigned to them, and Rumelhart and McClelland interpret those strength values as being related roughly monotonically to the likelihood the model would output those candidates. Since there is noise in some of the processes that contribute to strength values, they chose a threshold value (.2 on the 0–1 scale) and if a word surpassed that criterion, it was construed as being one of the model's guesses for the past tense form for a given stem. By this criterion, 24 of the 72 probe stems resulted in a strong tendency to incorrect responses—33% of the sample. Of these, 6 (*jump, pump, soak,*

[20] See Armstrong, Gleitman, and Gleitman (1983) for an analogous argument applied to conceptual categories.

warm, trail, glare) had no response at threshold. Though it is hard to reconstruct the reasons for this, two facts are worth noting. First, these verbs have no special resemblance to the apparently quasi-productive strong verb types—the factor that affects human responses. Second, the no-response verbs tend to cluster in phonetic similarity space either with one another (*jump, pump*) or with other verbs that the model erred on, discussed below (*soak/smoke; trail/mail; glare/tour*). This suggests that the reason for the model's muteness is that it failed to learn the relevant transformations; i.e. to generalize appropriately about the regular past. Apparently the steps taken to prevent the model from bogging down in insufficiently general case-by-case learning, such as blurring the Wickelfeatures and using noisy probabilistic output units during learning, did not work well enough.

But it also reveals one of the *inherent* deficits of the model we have alluded to: there is no such thing as a *variable* for *any* stem, regardless of its phonetic composition, and hence no way for the model to attain the knowledge that you can add /d/ to a "stem" to get its past. Rather, all the knowledge of the model consists of responses trained to the *concrete features in the training set*. If the new verbs happen not to share enough of these features with the words in the training set, or happen to possess features to which competing and mutually incompatible outputs had been associated, the model can fail to output any response significantly stronger than the background noise. The regular rule in symbolic accounts, in contrast, doesn't care what's in the word or how often its contents were submitted previously for training; the concept of a stem itself is sufficient. We return to this point when discussing some of the limitations of connectionist architecture in general.

Of the remaining 18 verbs for which the model did not output a single correct choice, 4 yielded grossly bizarre candidates:

(28) a. squat - squakt
b. mail - membled
c. tour - toureder
d. mate - maded

Three other candidates were far off the systematic mark:

(29) a. hug - hug
b. smoke - smoke
c. brown - brawned

Seven more showed a strong or exclusive tendency to double marking with the regular past tense morpheme (later we examine whether children make errors of this sort):

(30) a. type - typeded
b. step - steppeded
c. snap - snappeded
d. map - mappeded
e. drip - drippeded
f. carp - carpeded
g. smoke - smokeded

Note that the model shows an interesting tendency to make ill-advised vowel changes:

(31) a. shape - shipt
b. sip - sept
c. slip - slept
d. brown - brawned
e. mail - membled

Well before it has mastered the richly exemplified regular rule, the pattern-associator appears to have gained considerable confidence in certain incorrectly-grasped, sparsely exemplified patterns of feature-change among the vowels. This implies that a major "induction problem"—latching onto the productive patterns and bypassing the spurious ones—is not being solved successfully.

In sum, for 14 of the 18 stems yielding incorrect forms, the forms were quite removed from the confusions we might expect people to make. Taking these with the 6 no-shows, we have 20 out of the 72 test stems resulting in seriously wrong forms, a 28% failure rate. This is the state of the model after it has been trained 190–200 times on each item in a vocabulary of 336 regular verbs.

What we have here is not a model of the mature system.

6. On some common objections to arguments based on linguistic evidence

We have found that many psychologists and computer scientists feel uncomfortable about evidence of the sort we have discussed so far, concerning the ability of a model to attain the complex organization of a linguistic system in its mature state, and attempt to dismiss it for a variety of reasons. We consider the evidence crucial and decisive, and in this section we reproduce some of the objections we have heard and show why they are groundless.

"Those philosophical arguments are interesting, but it's really the empirical data that are important." All of the evidence we have discussed is empirical.

It is entirely conceivable that people could go around saying *What'z the answer?* or *He high-stuck the goalie* or *The canary pept* or *I don't know how she bore him* or *Yesterday we chat for an hour*, or that upon hearing such sentences, people could perceive them as sounding perfectly normal. In every case it is an empirical datum about the human brain that they don't. Any theory of the psychology of language must account for such data.

"Rule-governed behaviors indeed exist, but they are the products of schooling or explicit instruction, and are deployed by people only when in a conscious, reflective, problem-solving mode of thought that is distinct from the intuitive processes that PDP models account for" (see, for example, Smolensky, in press). This is completely wrong. The rule adding /d/ to a stem to form the past is not generally taught in school (it doesn't have to be!) except possibly as a rule of spelling, which if anything obscures its nature: for one thing, the plural morpheme, which is virtually identical to the past morpheme in its phonological behavior, is spelled differently ("s" versus "ed"). The more abstract principles we have discussed, such as distinctions between morphology and phonology, the role of roots in morphology, preservation of stem and affix identity, phonological processes that are oblivious to morphological origin, disjoint conditions for the application of morphological and phonological changes, distinct past tenses for homophones, interactions between the strong and regular systems, and so on, are consciously inaccessible and not to be found in descriptive grammars or language curricula. Many have only recently been adequately characterized; traditional prescriptive grammars tend to be oblivious to them or to treat them in a ham-fisted manner. For example, H.L. Mencken (1936) noted that people started to use the forms *broadcasted* and *joy-rided* in the 1920s (without consciously knowing it, they were adhering to the principle that irregularity is a property of verb roots, hence verbs formed from nouns are regular). The prescriptive guardians of the language made a fruitless attempt to instruct people explicitly to use *broadcast* and *joy-rode* instead, based on its similarity to *cast-cast* and *ride-rode*.

In fact, the objection gets the facts exactly backwards. One of the phenomena that the RM model is good at handling is unsystematic analogy formation based on its input history with subregular forms (as opposed to the automatic application of the regular rule where linguistically mandated). The irregular system, we have noted, is closely tied to memory as well as to language, so it turns out that people often have metalinguistic awareness of some of its patterns, especially since competing regular and irregular past tense forms carry different degrees of prestige and other socioeconomic connotations. Thus some of the fine points of use of the irregulars depend on exposure to standard dialects, on normative instruction, and on conscious reflection. Thus people, when in a reflective, conscious, problem-solving mode, will seem to act more like the RM model: the overapplication of subregularities that the model is prone to can be seen in modes of language use that bear all the hallmarks of self-conscious speech, such as jocularity (e.g. *spaghettus, I got schrod at Legal Seafood, The bear shat in the woods*), explicit instruction within a community of specialists (e.g. *VAXen* as the plural of *VAX*), pseudoerudition (*rhinoceri, axia* for *axioms*), and hypercorrection such as the anti-*broadcasted* campaign documented by Mencken (similarly, we found that some of our informants offered *Hurst no-hitted the Blue Jays* as their first guess as to the relevant past form but withdrew it in favor of *no-hit* which they "conceded" was "more proper").

"We academics speak in complex ways, but if you were to go down to [name of nearest working-class neighborhood] *you'd find that people talk very differently."* If anything is universal about language, it is probably people's tendency to denigrate the dialects of other ethnic or socioeconomic groups. One would hope that this prejudice is not taken seriously as a scientific argument; it has no basis in fact. The set of verbs that are irregular varies according to regional and socioeconomic dialect (see Mencken, 1936, for extensive lists), as does the character of the subregular patterns, but the principles organizing the system as a whole show no variation across classes or groups.

"Grammars may characterize some aspects of the ideal behavior of adults, but connectionist models are more consistent with the sloppiness found in children's speech and adult's speech errors, which are more 'psychological' phenomena." Putting aside until the next section the question of whether connectionist models really do provide a superior account of adult's or children's errors, it is important to recognize a crucial methodological asymmetry that this kind of objection fails to acknowledge. The ability to account for patterns of error is a useful criterion for evaluating competing theories *each of which can account for successful performance equally well*. But a theory that can *only* account for errorful or immature performance, with no account of why the errors are errors or how children mature into adults, is of limited value (Pinker, 1979, 1984; Wexler & Culicover, 1980; Gleitman & Wanner, 1982). (Imagine a "model" of the internal combustion engine that could mimic its ability to fail to start on cold mornings—by doing nothing—but could not mimic its ability to run, under any circumstances.)

Thus it is not legitimate to suggest, as Rumelhart and McClelland do, that "people—or at least children, even in early grade-school years—are not perfect rule-applying machines either. ... Thus we see little reason to believe that our model's 'deficiencies' are significantly greater than those of native speakers of comparable experience" (PDPII, p. 265–266). Unlike the RM model, no adult speaker is utterly stumped in an unpressured naturalistic situation when he or she needs to produce the past tense form of *soak* or *glare*, none vacillates between *kid* and *kidded*, none produces *membled* for *mailed* or *toureder* for *toured*. Although children equivocate in experimental tasks eliciting inflected nonce forms, these tasks are notorious for the degree to which they underestimate competence with the relevant phenomena (Levy, 1983; Maratsos et al., 1987; Pinker, Lebeaux, & Frost, 1987)—not to mention the fact that children do not remain children forever. The crucial point is that adults can speak without error and can realize that their errors are errors (by which we mean, needless to say, from the standpoint of the untaxed operation of their own system, not of a normative standard dialect). And children's learning culminates in adult knowledge. These are facts that any theory must account for.

7. The RM model and the facts of children's development

Rumelhart and McClelland stress that their model's ability to explain the developmental sequence of children's mastery of the past tense is the key point in favor of their model over traditional accounts. In particular, these facts are the "fine structure of the phenomena of language use and language acquisition" that their model is said to provide an exact account of, as opposed the traditional explanations which "leave out a great deal of detail", describing the phenomena only "approximately".

One immediate problem in assessing this claim is that there is no equally explicit model incorporating rules against which we can compare the RM model. Linguistic theories make no commitment as to how rules increase or decrease in relative strength during acquisition; this would have to be supplied by a learning mechanism that meshed with the assumptions about the representation of the rules. And theories discussed in the traditional literature of developmental psycholinguistics are far too vague and informal to yield the kinds of predictions that the RM model makes. There do exist explicit models of the acquisition of inflection, such as that outlined by Pinker (1984), but they tend to be complementary in scope to the RM model; the Pinker model, for example, attempts to account for how the child realizes that one word is the past tense version of another, and which of two competing past tense candidates is to be retained, which in the RM model is handled by the "teacher" or not at all, and relegates to a black box the process of abstracting the morphological and phonological changes relating past forms and stems, which is what the RM model is designed to learn.

The precision of the RM theory is surely a point in its favor, but it is still difficult to evaluate, for it is not obvious what features of the model give it its empirical successes. More important, it is not clear whether such features are consequences of the model's PDP architecture or simply attributes of fleshed-out processes that would function in the same way in any equally-explicit model of the acquisition process. In most cases Rumelhart and McClelland do not apportion credit or blame for the model's behavior to specific aspects of its operation; the model's output is compared against the data rather globally. In other cases the intelligence of the model is so distributed and its output mechanisms are so interactive that it is difficult for anyone to know what aspect of the model makes it successful. And in general, Rumelhart and McClelland do not present critical tests between competing hypotheses embodying minimally different assumptions, only descriptions of goodness of fit between their model and the data. In this section, we unpack the assumptions of the model, and show which ones are doing the work in accounting for the developmental facts—and whether the developmental facts are accounted for to begin with.

7.1. Unique and shared properties of networks and rule systems

Among the RM model's many properties, there are two that are crucial to its accounts of developmental phenomena. First, it has a learning mechanism that makes it *type-frequency sensitive*: the more verbs it encounters that embody a given type of morphophonological change, the stronger are its graded representations of that morphophonological change, and the greater is the tendency of the model to generalize that change to new input verbs. Furthermore, the different past tense versions of a word that would result from applying various regularities to it are computed in parallel and there is a *competition* among them for expression, whose outcome is determined mainly by the strength of the regularity and the goodness of the match between the regularity and the input. (In fact the outcome can also be a blend of competing responses, but the issue of response blending is complex enough for us to defer discussing it to a later section.)

It is crucial to realize that neither frequency-sensitivity nor competition is unique to PDP models. Internal representations that have graded strength values associated with them are probably as old as theories of learning in psychology; in particular, it is commonplace to have greater strength values assigned to representations that are more frequently exemplified in the input during learning, so that strength of a representation basically corresponds to degree of confidence in the hypothesis represented. Competition among candidate operations that partially match the input is also a ubiquitous assumption among symbol-processing models in linguistics and cognitive psychology. Spreading-activation models and production systems, which are prototypical symbol-processing models of cognition, are the clearest examples (see, e.g. Newell & Simon, 1972; Anderson, 1976, 1983; MacWhinney & Sokolov, 1987).

To show how these assumptions are part and parcel of standard rule-processing models, we will outline a simplified module for certain aspects of past tense acquisition, which searches for the correct past tense rule or rules, keeping several candidates as possibilities before it is done. We do not mean to propose it as a serious theory, but only as a demonstration that many of the empirical successes of the RM model are the result of assumptions about frequency-sensitivity and competition among output candidates that are independent of parallel distributed processing in networks of simple units.

A simple illustrative module of a rule-based inflection acquisition theory, incorporating assumptions about frequency-sensitivity and competition

Acquiring inflectional systems poses a number of tricky induction problems, discussed at length in Pinker (1984). When a child hears an inflected verb in a single context, it is utterly ambiguous what morphological category the inflection is signaling (the gender, number, person, or some combination of those agreement features for the subject? for the object? is it tense? aspect? modality? some combination of these?). Pinker (1984) suggested that the child solves this problem by "sampling" from the space of possible hypotheses defined by combinations of an innate finite set of elements, maintaining these hypotheses in the provisional grammar, and testing them against future uses of that inflection, expunging a hypothesis if it is counterexemplified by a future word. Eventually, all incorrect hypotheses about the category features encoded by that affix will be pruned, any correct one will be hypothesized, and only correct ones will survive.

The surviving features define the dimensions of a word-specific paradigm structure into whose cells the different inflected forms of a given verb are placed (for example, singular–plural or present–past–future). The system then seeks to form a productive general paradigm—that is, a set of rules for related inflections—by examining the patterns exhibited across the paradigms for the individual words. This poses a new induction problem because of the large number of possible generalizations consistent with the data, and it cannot be solved by examining a single word-specific paradigm or even a set of paradigms. For example, in examining *sleep/slept*, should one conclude that the regular rule of English laxes and lowers the vowel and adds a *t*? If so, does it do so for all stems or only for those ending in a stop, or only those whose stem vowel is *i*? Or is this simply an isolated irregular form, to be recorded individually with no contribution to the regular rule system? There is no way to solve the problem other than by trying out various hypotheses and seeing which ones survive when tested against the ever-growing vocabulary. Note that this induction problem is inherent to the task and cannot be escaped from using connectionist mechanisms or any other mechanisms; the RM model attempts to solve the problem in one way, by trying out a large number of hypotheses of a certain type in parallel.

A symbolic model would solve the problem using a mechanism that can formulate, provisionally maintain, test, and selectively expunge hypotheses about rules of various degrees of generality. It is this hypothesis-formation mechanism that the simplified module embodies. The module is based on five assumptions:

1. Candidates for rules are hypothesized by comparing base and past tense versions of a word, and factoring apart the changing portion, which serves as the rule operation, from certain morphologically-relevant phonological components of the stem, which serve to define the class of stems over which the operation can apply.[21] Specifically, let us assume that when the addition of material to the edge of a base form is noted, the added material is stored as an affix, and the provisional definition of the morphological class will consist of the features of the edge of the stem to which the affix is attached. When a vowel is noted to change, the change is recorded, and the applicable morphological class will be provisionally defined in terms of the features of the adjacent consonants. (In a more realistic model, global properties defining the "basic words" of a language, such as monosyllabicity in English, would also be extracted.)
2. If two rule candidates have been coined that have the same change operation, a single collapsed version is created, in which the phonological features distinguishing their class definitions are eliminated.
3. Rule candidates increase in strength each time they have been exemplified by an input pair.
4. When an input stem has to be processed by the system in its intermediate stages, an input is matched in parallel against all existing rule candidates, and if it falls into several classes, several past tense forms may be generated.
5. The outcome of a competition among the past tense forms is determined by the strength of the relevant rule and the proportion of a word's features that were matched by that rule.

[21] More accurately, the changing portion is examined subsequent to the subtraction of any phonological and phonetic changes that have been independently acquired.

The model works as follows. Imagine its first input pair is *speak/spoke*. The changing portion is *i* → *o*. The provisional definition of the class to which such a rule would apply would be the features of the adjacent consonants, which we will abbreviate as *p_k*. Thus the candidate rule coined is (32a), which can be glossed as "change *i* to *o* for the class of words containing the features of /p/ before the vowel and containing the features of /k/ after the vowel". Of course, the candidate rule has such a specific class definition in the example that it is almost like listing the pair directly. Let us make the minimal assumptions about the strength function, and simply increase it by 1 every time a rule is exemplified. Thus the strength of this rule candidate is 1. Say the second input is *get/got*. The resulting rule candidate, with a strength of 1, is (32b). A regular input pair, *tip/tipped*, would yield (32c). Similarly, *sing/sang* would lead to (32d), and *hit/hit* would lead to (32e), each with unit strength,

(32) a. Change: i → o
Class: p_k
b. Change: e → ɔ
Class: g_t
c. Suffix: t
Class: p#
d. Change: i → æ
Class: s_ŋ
e. Suffix: ∅
Class: t#
Change: i → i
Class: h_t.[22]

Now we can examine the rule-collapsing process. A second regular input, *walk/walked*, would inspire the learner to coin the rule candidate (33a) which, because it shares the change operation of rule candidate (32c), would be collapsed with it to form a new rule (33b) of strength 2 (summing the strengths of its contributing rules, or equivalently, the number of times it has been exemplified).

[22] Let us assume that it is unclear to the child at this point whether there is a null vowel change or a null affix, so both are stored. Actually, we don't think either is accurate, but it will do for the present example.

(33) a. Suffix: t
Class: k#

b. Suffix: t
Class: C #
[-voiced]
[-continuant]
[-sonorant]

The context-collapsing operation has left the symbol "C" (for consonant) and its three phonological features as the common material in the definitions of the two previously distinct provisional classes.

Now consider the results of a third regular input, *pace/paced*. First, a fairly word-specific rule (34a) would be coined; then it would be collapsed with the existing rule (33) with which it shares a change operation, yielding a rule (34b) with strength 3.

(34) a. Suffix: t
Class: s#

b. Suffix; t
Class: C #
[-voiced]

Rule candidates based on subregularities would also benefit from the increases in strength that would result from the multiple input types exemplifying it. For example, when the pair *ring/rang* is processed, it would contribute (35a), which would then be collapsed with (32d) to form (35b). Similar collapsing would strengthen other subregularities as tentative rule candidates, such as the null affix.

(35) a. Change: i → a
Class: r_ŋ
b. Change: i → a
Class: C_ŋ

Though this model is ridiculously simple, one can immediately see that it has several things in common with the RM model. First, regularities, certain subregularities, and irregular alternations are extracted, to be entertained as possible rules, by the same mechanism. Second, mechanisms embodying the different regularities accrue strength values that are monotonically related to the number of inputs that exemplify them. Third, the model can generalize to new inputs that resemble those it has encountered in the past; for example, *tick*, which terminates in an unvoiced stop, matches the context of rule (34b), and the rule can add a /t/ to the end of it as a result to form *ticked*. Fourth, a new input can match several rules at the same time. For example, *bet* will match one rule candidate because it ends in an unvoiced stop and it will match another because it ends in *t*. The exact strengths of the competing alternatives will depend on the strengths of the candidate rules and on the goodness of match between the stem and the class definitions associated with the rules.[23]

This candidate-hypothesization module can only be part of the mechanism that acquires the past tense system. Other mechanisms or principles, such as those discussed in Pinker (1984), must evaluate the rule candidates and eliminate the incorrect ones, such as those that simply characterize lists of similar strong forms, and must retain any genuine rules in a general paradigm. As noted in Section 4.4, regular rules are distinguished by applying in the default or "elsewhere" case. One can imagine the following learning strategy, which can be called the "Nonexceptional Exceptions to Exceptions Strategy", that would discover regular rules using this criterion. Comparing the acquired stem-past pairs whose first member contains *eep*, the child would notice that there are many exceptions to the tentative *eep* → *ept* rule candidate and that most of the exceptions to it themselves follow the pattern holding of verbs whose present forms do not contain *eep* (*seeped, peeped, steeped, beeped*, etc.). Furthermore, exceptions to other subregularities such as *bend/bent–lend/lent* will also largely obey the pattern holding of verbs lacking *end* (*end/ended, fend/fended, mend/mended*, etc.). Thus, within the child's lexicon one regularity, the addition of /d/, knows no phonological bounds, and can potentially apply to any base form, whereas this is not true of any other regularity. In this way, some regularities can be enshrined as permanent productive rules whereas others can be discarded or treated differently.

Other constraints contributed by other principles and components of grammar would also influence the extraction and sorting of putative rules. For example, the syntax and lexicon would segregate derived forms out of these

[23] Note also that the strongest output among competing candidates for the past form of a given verb could change as a function of the input history of the model. For example, during the first five inputs the only output for *speak* would be its irregular past *spoke*. After the sixth input, the regularized past version *speaked* would also be provided, by rule (34b), though the strength of this output would be low because the regular rule would not be strong enough to overcome the strength of the irregular form resulting from its very close match to (32a). If candidate strength is equal to [proportion of stem features matched × strength of matching rule], the irregular output would have a strength of (.75 × 1) = .75, whereas the regular rule would have, say (.2 × 2) = .4 (the exact numbers are not crucial here). However, after a number of inputs, the regular rule has increased in strength to 4, and so the strength of *speaked* would be (.2 × 4) = .8, making it stronger than the irregular form *hit*. In this way, a rule-finding module could overgeneralize in its intermediate stages, erring on verbs that it previously handled properly, for similar reasons that the phenomenon occurs in the RM model. Later we examine whether this is the correct explanation for children's behavior.

calculations, and the phonology would subtract out modifications abstracted from consonant clusters of simple words and perhaps from sets of morphologically unrelated rules. Finally, the general regular paradigm would be used when needed to fill out empty cells of word-specific paradigms with a unique entry, while following the constraint that irregular forms in memory block the product of the regular rule, and only a single form can be generated for a specific stem when more than one productive rule applies to it (multiple entries can exist only when the irregular form is too weakly represented, or when both multiple forms are witnessed in the input; see Pinker, 1984).

Though both our candidate-hypothesization module and the RM model share certain properties, let us be clear about the differences. The RM model is designed to account for the entire process that maps stems to past tense forms, with no interpretable subcomponents, and few constraints on the regularities that can be recorded. The candidate-hypothesization module, on the other hand, is meant to be a part of a larger system, and its outputs, namely rule candidates, are symbolic structures that can be examined, modified or filtered out by other components of grammar. For example, the phonological acquisition mechanism can note the similarities between t/d/id and s/z/iz and pull out the common phonological regularities, which would be impossible if those allomorphic regularities were distributed across a set of connection weights onto which countless other regularities were superimposed.

It is also important to note that, as we have mentioned, the candidate-hypothesization module is motivated by a requirement of the learnability task facing the child. Specifically, the child at birth does not know whether English has a regular rule, or if it does, what it is or whether it has one or several. He or she must examine the input evidence, consisting of pairs of present and past forms acquired individually, to decide. But the evidence is locally ambiguous in that the nonproductive exceptions to the regular rule are not a random set but display some regularities for historical reasons (such as multiple borrowings from other languages or dialects, or rules that have ceased to be productive) and psychological reasons (easily-memorized forms fall into family resemblance structures). So the child must distinguish real from apparent regularities. Furthermore, there is the intermediate case presented by languages that have several productive rules applying to different classes of stems. The "learnability problem" for the child is to distinguish these cases. Before succeeding, the child must entertain a number of candidates for the regular rule or rules, because it is only by examining large sets of present-past pairs that the spurious regularities can be ruled out and the partially-productive ones assigned to their proper domains; small samples are always ambiguous in this regard. Thus a child who has not yet solved the problem of distinguishing general productive rules from restricted productive rules from accidental patterns will have a number of candidate regularities still open as hypotheses. At this stage there will be competing options for the pas⁺ tense form of a given verb. The child who *has not yet figured out* the distinction between regular, subregular, and idiosyncratic cases will display behavior that is similar to a system that is *incapable of making* the distinction—the RM model.

In sum, any adequate rule-based theory will have to contain a module that extracts multiple regularities at several levels of generality, assign them strengths related to their frequency of exemplification by input verbs, and let them compete in generating a past tense form for a given verb. In addition, such a model can attain the adult state by feeding its candidates into paradigm-organization processes, which, following linguistic constraints, distinguish real generalizations from spurious ones. With this alternative model in mind, we can now examine which aspects of the developmental data are attributable to specific features of the RM model's parallel distributed processing architecture—specifically, to its collapsing of linguistic distinctions—and those which are attributable to its assumptions of graded strength, type-frequency sensitivity, and competition which it shares with symbolic alternatives.

7.2. Developmental phenomena claimed to support the Rumelhart–McClelland model

The RM model is, as the authors point out, very rich in its empirical predictions. It is a strong point of their model that it provides accounts for several independent phenomena, all but one of them unanticipated when the model was designed. They consider four phenomena in detail: (1) the U-shaped curve representing the overregularization of strong verbs whose regular pasts the child had previously used properly; (2) The fact that verbs ending in *t* or *d* (e.g. *hit*) are regularized less often than other verbs; (3) The order of acquisition of the different classes of irregular verbs manifesting different subregularities; (4) The appearance during the course of development of [past + *ed*] errors such as *ated* in addition to [stem + *ed*] errors such as *eated*.

7.2.1. Developmental sequence of productive inflection (the "U"-shaped curve)

It is by now well-documented that children pass through two stages before attaining adult competence in handling the past tense in English. In the first stage, they use a variety of correct past tense forms, both irregular and regular, and do not readily apply the regular past tense morpheme to nonce words presented in experimental situations. In the second stage, they apply the past tense morpheme productively to irregular verbs, yielding overregularizations such as *hitted* and *breaked* for verbs that they may have used exclusively in their correct forms during the earlier stage. Correct and overregularized forms coexist for an extended period of time in this stage, and at some point during that stage, children demonstrate the ability to apply inflections to nonce forms in experimental settings. Gradually, irregular past tense forms that the child continues to hear in the input drive out the overregularized forms he or she has created productively, resulting in the adult state where a productive rule coexists with exceptions (see Berko, 1958; Brown, 1973; Cazden, 1968; Ervin, 1964; Kuczaj, 1977, 1981).

A standard account of this sequence is that in the first stage, with no knowledge of the distinction between present and past forms, and no knowledge of what the regularities are in the adult language that relate them, the child is simply memorizing present and past tense forms directly from the input. He or she correctly uses irregular forms because the overregularized forms do not appear in the input and there is no productive rule yet. Regular past tenses are acquired in the same way, with no analysis of them into a stem plus an inflection. Using mechanisms such as those sketched in the preceding section, the child builds a productive rule and can apply it to any stem, including stems of irregular verbs. Because the child will have had the opportunity to memorize irregular pasts before relating stems to their corresponding pasts and before the evidence for the regular relationship between the two has accumulated across inputs, correct usage can in many cases precede overregularization. The adult state results from a realization, which may occur at different times for different verbs, that overregularized and irregular forms are both past tense versions of a given stem, and by the application of a Uniqueness principle that, roughly, allows the cells of an inflectional paradigm for a given verb to be filled by no more and no less than one entry, which is the entry witnessed in the input if there are competing nonwitnessed rule-generated forms and witnessed irregulars (see Pinker, 1984).

The RM model also has the ability to produce an arbitrary past tense form for a given present when they have been exemplified in the input, and to generate regular past tense forms for the same verbs by adding *-ed*. Of course, it does so without distinct mechanisms of rote and rule. In early stages, the links between the Wickelfeatures of a base irregular form and the Wickelfeatures of its past form are given higher weights. However, as a diverse set of regular forms begins to stream in, links are strengthened between a large set of input Wickelfeatures and the output Wickelfeatures containing features of the regular past morpheme, enough to make the regularized form a stronger output than the irregular form. During the overregularization stage, "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" (PDPII, p. 268). The irregular form eventually returns as the strongest output because repeated presentations of it cause the network to tune the connection weights so that the Wickelfeatures that are specific to the irregular stem form (and to similar irregular forms manifesting the same kind of stem-past variation) are linked more and more strongly to the Wickelfeatures specific to their past forms, and develop strong negative weights to the Wickelfeatures corresponding to the regular morpheme. That is, the prevalence of a general pattern across a large set of verbs trades off against the repeated presentation of a single specific pattern of a single verb presented many times (with subregularities constituting an intermediate case). This gives the model the ability to be either conservative (correct for an irregular verb) or productive (overregularizing an irregular verb) for a given stem, depending on the mixture of inputs it has received up to a given point.

Since the model's tendency to generalize lies on a continuum, any sequence of stages of correct irregulars or overregularized irregulars is possible in principle, depending on the model's input history. How, then, is the specific shift shown by children, from correct irregular forms to a combination of overregularized and correct forms, mimicked by the model? Rumelhart and McClelland divide the training sequence presented to the model into two stages. In the first, they presented 10 high-frequency verbs to the model, 2 of them regular, 10 times each. In the second, they added 410 verbs to this sample, 334 of them regular, and presented the sample of 420 verbs 190 times. The beginning of the downward arm of the U-shaped plot of percent correct versus time, representing a worsening of performance for the irregular verbs, occurs exactly at the boundary between the first set of inputs and the second. The sudden influx of regular forms causes the links capturing the regular pattern to increase in strength; prior to this influx, the regular pattern was exemplified by only two input forms, not many more than those exemplifying any of the idiosyncratic or subregular patterns. The shift from the first to the second stage of the model's behavior, then, is a direct consequence of a shift in the input mixture from a heterogeneous collection of patterns to a collection in which the regular pattern occurs in the majority.

It is important to realize the theoretical claim inherent in this demonstration. *The model's shift from correct to overregularized forms does not emerge from any endogenous process; it is driven directly by shifts in the environment.* Given a different environment (say, one in which heterogeneous irregular forms suddenly start to outnumber regular forms), it appears that the model could just as easily go in the opposite direction, regularizing in its first stage and then becoming accurate with the irregular forms. In fact, since the model always has the potential to be conservative or rule-governed, and continuously tunes itself to the input, it appears that just about any shape of curve at all is possible, given the right shifts in the mixture of regular and irregular forms in the input.

Thus if the model is to serve as a theory of children's language acquisition, Rumelhart and McClelland must attribute children's transition between the first and second stage to a prior transition of the mixture of regular and irregular inputs from the external environment. They conjecture that such a transition might occur because irregular verbs tend to be high in frequency. "Our conception of the nature of [the child's] 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" (p. 241). They concede that there is no abrupt shift in the input to the child, but suggest that children's acquisition of the present tense forms of verbs serves as a kind of filter for the past tense learning mechanism, and that this acquisition of base forms undergoes an explosive growth at a certain stage of development. Because the newly-acquired verbs are numerous and presumably lower in frequency than the small set of early-acquired verbs, it will include a much higher proportion of regular verbs. Thus the shift in the proportion of regular verbs in the input to the model comes about as a consequence of a shift from high frequency to medium frequency verbs; Rumelhart and McClelland do not have to adjust the leanness or richness of the input mixture by hand.

The shift in the model's input thus is not entirely ad hoc, but is it realistic? The use of frequency counts of verbs in written samples in order to model children's vocabulary development is, of course, tenuous.[24] To determine whether the input to children's past tense learning shifts in the manner assumed by Rumelhart and McClelland, we examined Roger Brown's unpublished grammars summarizing samples of 713 utterances of the spontaneous speech of three children observed at five stages of development. The stages were defined in terms of equally spaced intervals of the children's Mean Length of Utterance (MLU). Each grammar includes an exhaustive list of the child's verbs in the sample, and an explicit discussion of whether the child

[24] For example, in the Kucera and Francis (1967) counts used by Rumelhart and McClelland, medium frequencies are assigned to the verbs *flee*, *seek*, *mislead* and *arise*, which are going to be absent from a young child's vocabulary. On the other hand *stick* and *tear*, which play a significant role in the ecology of early childhood, are ranked as low-frequency. *Be* and *do* are not in the high-frequency group, where they belong—*do* belongs because of its ubiquity in questions, a fact not reflected in the written language. *Be* appears to be out of the study, perhaps because Rumelhart and McClelland count the frequency of the *-ing* forms.

was overregularizing the past tense rule.[25] In addition, we examined the vocabulary of Lisa, the subject of a longitudinal language acquisition study at Brandeis University, in her one-word stage. Two of the children, Adam and Eve, began to overregularize in the Stage III sample; the third child, Sarah, began to overregularize only in the State V sample except for the single form *heared* appearing in Stage II which Brown noted might simply be one of Sarah's many cases of unusual pronunciations. We tabulated the size of each child's verb vocabulary and the proportion of verbs that were regular at each stage.[26]

The results, shown in Table 1 and Figure 2, are revealing. The percentage of the children's verbs that are regular is remarkably stable across children and across stages, never veering very far from 50%. (This is also true in parental speech itself: Slobin, 1971, showed that the percentage of regular verbs in Eve's parents' speech during the period in which she was overregularizing was 43%.) In particular, there is no hint of a consistent increase in the proportion of regular verbs prior to or in the stage at which regularizations first occur. Note also that an explosive growth in vocabulary does not invariably precede the onset of regularization. This stands in stark contrast to the assumed input to the RM model, where the onset of overregularization occurs subsequent to a sudden shift in the proportion of regular forms in the input from 20% to 80%. Neither the extreme rarity of regular forms during the conservative stage, nor the extreme prevalence of regular forms during the overproductive stage, nor the sudden transition from one input mixture to another, can be seen in human children. The explanation for their developmental sequence must lie elsewhere.

We expect that this phenomenon is quite general. The plural in English, for example, is overwhelmingly regular even among high-frequency nouns:[27] only 4 out of the 25 most frequent concrete count nouns in the Francis and Kucera (1982) corpus are irregular. Since there are so few irregular plurals, children are never in a stage in which irregulars strongly outnumber regulars

[25] For details of the study, see Brown (1973); for descriptions of the unpublished grammars, see Brown (1973) and Pinker (1984). Verification of some of the details reported in the grammars, and additional analyses of children's speech to be reported in this paper, were based on on-line transcripts of the speech of the Brown children included in the Child Language Data Exchange System; MacWhinney & Snow (1985).

[26] A verb was counted whether it appeared in the present, progressive, or past tense form, and was counted only once if it appeared in more than one form. Since most of the verbs were in the present, this is of little consequence. We counted a verb once across its appearances alone and with various particles since past tense inflection is independent of these differences. We excluded modal pairs such as *can/could* since they only occasionally encode a present/past contrast for adults. We excluded catenative verbs that encode tense and mood in English and hence which do not have obvious past tenses such as in *going to*, *come on*, and *gimme*.

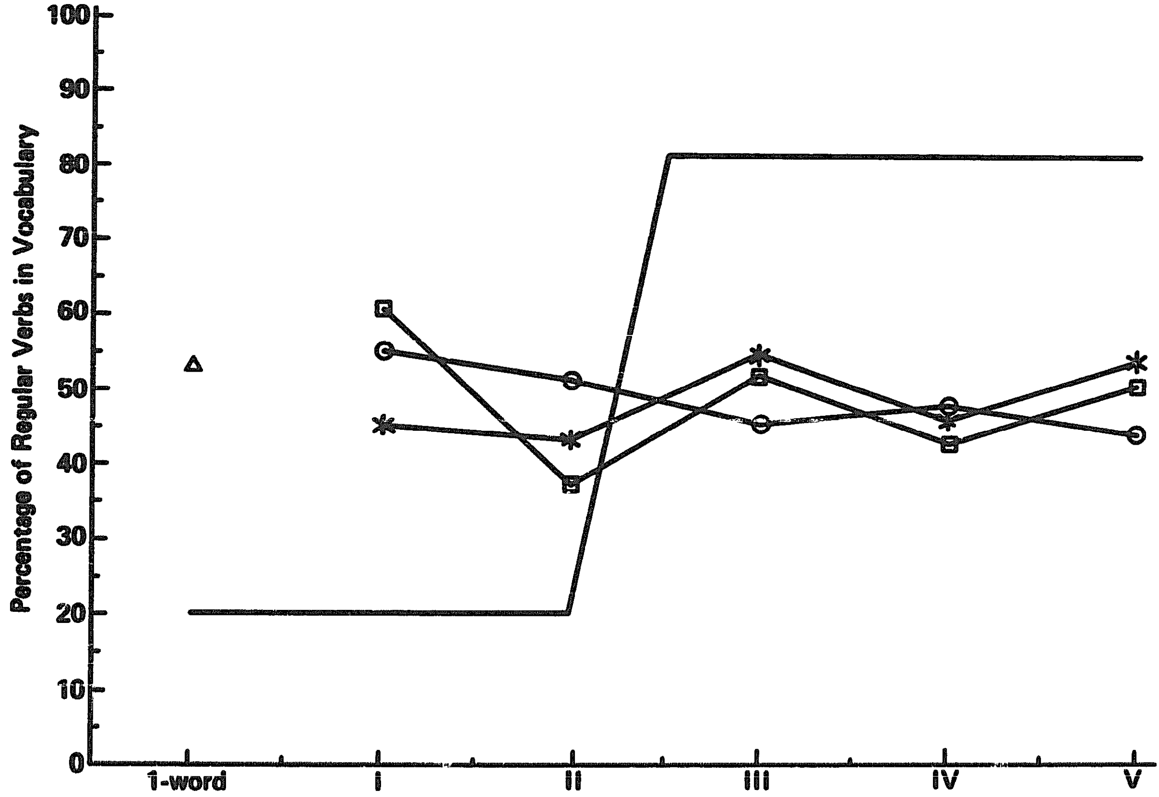
[27] We are grateful to Maryellen McDonald for this point.

Table 1. *Proportion of children's verbs that have regular past tense forms*

	Stage					
	1-Word	I	II	III	IV	V
Adam	–	.45(31)	.43(44)*	.55(83)	.46(83)	.54(78)
Eve	–	.55(31)	.51(49)*	.45(53)	.48(58)	.44(45)
Sarah	–	.61(18)	.37(49)	.52(44)	.43(58)	.51(84)*
Lisa	.53(53)	–	–	–	–	–
Mean for Adam, Eve, & Sarah		.54	.44	.51	.46	.50

Note: Size of verb vocabulary is listed in parentheses. An asterisk indicates the stage at which the child began overregularizing.

Figure 2. *The percentage of verbs that are regular in four children's vocabularies at different stages (as defined by Brown, 1973). The predictions of the Rumelhart–McClelland model are shown for comparison purposes, under the assumption that regularization begins in Stage III. Key*: —*— Adam, —⊖— Eve, —⊟— Sarah, —△— Lisa, —— Rumelhart–McClelland Model.



in the input or in their vocabulary of noun stems. Nonetheless, the U-shaped developmental sequence can be observed in the development of plural inflection in the speech of the Brown children: for example, Adam said *feet* nine times in the samples starting at age 2;4 before he used *foots* for the first time at age 3;9; Sarah used *feet* 18 times starting at 2;9 before uttering *foots* at 5;1; Eve uttered *feet* a number of times but never *foots*.

Examining *token* frequencies only underlines the unnaturally favorable assumptions about the input used in the RM model's training run. Not only does the transition from conservatism to overregularization correspond to a shift from a 20/80 to an 80/20 ratio of regulars to irregulars, but in the first, conservative phase, high-frequency irregular pairs such as *go/went* and *make/made* were only presented 10 times each, whereas in the overregularizing phase the hundreds of regular verbs were presented 190 times each. In contrast, irregular verbs are always much higher in token frequency in children's environment. Slobin (1971) performed an exhaustive analysis of the verbs heard by Eve in 49 hours of adults' speech during the phase in which she was overregularizing and found that the ratio of irregular to regular tokens was 3:1. Similarly, in Brown's smaller samples, the ratios were 2.5:1 for Adam's parents, 5:1 for Eve's parents, and 3.5:1 for Sarah's parents. One wonders whether presenting the RM model with 10 high-frequency verbs, say, 190 times each in the first phase could have burned in the 8 irregulars so strongly that they would never be overregularized in Phase 2.

If children's transition from the first to the second phase is not driven by a change in their environments or in their vocabularies, what causes it? One possibility is that a core assumption of the RM model, that there is no psychological reality to the distinction between rule-generated and memorized forms, is mistaken. Children might have the capacity to memorize independent present and past forms from the beginning, but a second mechanism that coins and applies rules might not go into operation until some maturational change put it into place, or until the number of verbs exemplifying a rule exceeded a threshold. Naturally, this is not the only possible explanation. An alternative is that the juxtaposition mechanism that relates each stem to its corresponding past tense form has not yet succeeded in pairing up memorized stems and past forms in the child's initial stage. No learning of the past tense regularities has begun because there are no stem-past input pairs that can be fed into the learning mechanism; individually acquired independent forms are the only possibility.

Some of the evidence supports this alternative. Brown notes in the grammars that children frequently used the present tense form in contexts that clearly called for the past, and in one instance did the reverse. As the children developed, past tense forms were used when called for more often, and evidence for an understanding of the function of the past tense form and the tendency to overregularize both increase. Kuczaj (1977) provides more precise evidence from a cross-sectional study of 14 children. He concluded that once children begin to regularize they rarely use a present tense form of an irregular verb in contexts where a past is called for.

The general point is that in either case *the RM model does not explain children's developmental shift from conservatism to regularization. It attempts to do so only by making assumptions about extreme shifts in the input to rule learning that turn out to be false.* Either rules and stored forms are distinct, or some process other than extraction of morphophonological regularity explains the developmental shift. The process of coming to recognize that two forms constitute the present and past tense variants of the same verb, that is, the juxtaposition process, seems to be the most likely candidate.

Little needs to be said about the shift from the second stage, in which regularization and overregularization occurs, to the third (adult) stage, in which application of the regular rule and storage of irregular pasts cooccur. Though the model does overcome its tendency to overregularize previously acquired irregular verbs, we have shown in a previous section that it never properly attains the third stage. This stage is attained, we suggest, not by incremental strength changes in a pattern-finding mechanism, but by a mechanism that makes categorical decisions about whether a hypothesized rule candidate is a genuine productive rule and about whether to apply it to a given verb.

On the psychological reality of the memorized/rule-generated distinction. In discussing the developmental shift to regularization, we have shown that there can be developmental consequences of the conclusion that was forced upon us by the linguistic data, namely that rule-learning and memorization of individual forms are separate mechanisms. (In particular, we pointed out that one might mature before the other, or one requires prior learning—juxtaposing stems and past forms—and the other does not.) This illustrates a more general point: the psychological reality of the memorized/rule-generated distinction predicts the possibility of finding dissociations between the two processes, whereas a theory such as Rumelhart and McClelland's that denies that reality predicts that such dissociations should not be found. The developmental facts are clearly on the side of there being such a distinction.

First of all, children's behavior with irregular past forms during the first, pre-regularization phase bears all the signs of rote memorization, rather than a tentatively overspecific mapping from a specific set of stem features to a specific set of past features. Brown notes, for example, that Adam used *fell-down* 10 times in the Stage II sample without ever using *fall* or *falling*, so his production of *fell-down* cannot be attributed to any sort of mapping at all from stem to past. Moreover there is no hint in this phase of any interaction or transfer of learning across phonetically similar individual irregular forms: for example, in Sarah's speech, *break/broke* coexisted with *make/made* and neither had any influence on *take*, which lacked a past form of any sort in her speech over several stages. Similar patterns can be found in the other children's speech.

A clear example of a dissociation between rote and rule over a span in which they coexist comes from Kuczaj (1977), who showed that children's mastery of irregular past tense forms was best predicted by their chronological age, but their mastery of regular past tense forms was best predicted by their Mean Length of Utterance. Brown (1973) showed that MLU correlates highly with a variety of measures of grammatical sophistication in children acquiring English. Kuczaj's logic was that irregular pasts are simply memorized, so the sheer number of exposures, which increases as the child lives longer, is the crucial factor, whereas regular pasts can be formed by the application of a rule, which must be induced as part of the child's developing grammar, so overall grammatical development is a better predictor. Thus the linguistic distinction between lists of exceptions and rule-generated forms (see Section 4.4) is paralleled by a developmental distinction between opportunities for list-learning and sophistication of a rule system.

Another possible dissociation might be found in individual differences. A number of investigators of child language have noted that some children are conservative producers of memorized forms whereas others are far more willing to generalize productively. For example, Cazden (1968) notes that "Adam was more prone to overgeneralizations than Eve and Sarah" (p. 447), an observation also made by Brown in his unpublished grammars. More specifically, Table 1 shows that Sarah began to regularize the past tense two stages later than the other two children despite comparable verb vocabularies. Maratsos et al. (1987) documented many individual differences in the willingness of children to overgeneralize the causative alternation. If such differences do not reflect differences in the children's environments or vocabularies (they don't in the case of the past tense), presumably they result from the generalizing mechanism of some children being stronger or more developed than that of others, without comparable differences in their ability to record forms directly from the input. The RM model cannot easily account for any of these dissociations (other than by attributing crucial aspects of the generalization phenomena to mechanisms entirely outside their model), because memorized forms and generalizations are handled by a single mechanism—recall that the identity map in the network must be learned by adjusting a large set of connection weights, just like any of the stem alterations; it is not there at the outset, and is not intrinsically easy to learn.

The question is not closed, but the point is that the different theories can in principle be submitted to decisive empirical tests. It is such tests that should be the basis for debate on the psychological issue at hand. Simply demonstrating that there exist contrived environments in which a network model can be made to mimic some data, especially in the absence of comparisons to alternative models, tells us nothing about the psychology of the child.

7.2.2. Performance with no-change verbs

A class of English verbs does not change in form between stem and past: *beat, cut, put, hit*, and others. All of these verbs end in a *t* or *d*. Bybee and Slobin (1982) suggest that this is no coincidence. They suggest that learners generate a schema for the form of past tense verbs on the basis of prevalent regular forms which states that past tense verbs end in *t* or *d*. A verb whose stem already ends in *t* or *d* spuriously appears to have already been inflected for past tense, and the child is likely to assume that it *is* a past tense form. As a result, it can be entered as the past version of the verb in the child's paradigm, blocking the output of the regular rule. Presumably this tendency could result in the unchanged verb surviving into adulthood, causing the no-change verbs to have entered in the language at large in some past generation and to be easily relearned thereafter. We will call this phenomenon *misperception*.[28]

In support of this hypothesis, Bybee and Slobin found in an elicitation experiment that for verbs ending in *t* or *d*, children were more likely to produce a past tense form identical to the present than a regularized form, whereas for verbs not ending in a *t* or *d*, they were more likely to produce a regularized form than an unchanged form. In addition, Kuczaj (1978) found in a judgment task that children were more likely to accept correct no-change forms for nonchanging verbs than correct past tense forms for other irregular verbs such as *break* or *send*, and less likely to accept overregularized versions of no-change verbs than overregularized versions of other irregular verbs. Thus not only do children learn that verbs ending in *t/d* are likely to be unchanged, but this subregularity is easier for them to acquire than the kinds of changes such as the vowel alternations found in other classes of irregular verbs.

Unlike the three-stage developmental sequence for regularization, chil-

[28] Bybee and Slobin do not literally propose that the child mis*analyzes t/d*-final verbs as (nonexistent) stems inflected by a rule. Rather, they postulate a static template which the child matches against unanalyzed forms during word perception to decide whether the forms are in the past tense or not.

dren's sensitivity to the no-change subregularity for verbs ending in *t/d* played no role in the design of the RM model or of its simulation run. Nonetheless, Rumelhart and McClelland point out that during the phase in which the model was overregularizing, it produced stronger regularized past tense candidates for verbs not ending in *t/d* than for verbs ending in *t/d*, and stronger unchanged past candidates for verbs ending in *t/d* than for verbs not ending in *t/d*. This was true not only across the board, but also within the class of regular verbs, and within the classes of irregular verbs that do change in the past tense, for which no-change responses are incorrect. Furthermore, when Rumelhart and McClelland examined the total past tense response of the network (that is, the set of Wickelfeatures activated in the response pool) for verbs in the different irregular subclasses, they found that the no-change verbs resulted in fewer incorrectly activated Wickelfeatures than the other classes of irregulars. Thus both aspects of the acquisition of the no-change pattern fall out of the model with no extra assumptions.

Why does the model display this behavior? Because the results of its learning are distributed over hundreds of thousands of connection weights, it is hard to tell, and Rumelhart and McClelland do not try to tease apart the various possible causal factors. Misperception cannot be the explanation because the model always received correct stem-past pairs. There are two other possibilities. One is that connections from many Wickelfeatures to the Wickelfeatures for word-final *t*, and the thresholds for those Wickelfeatures, have been affected by the many *regular* stem-past pairs fed into the model. The response of the model is a blend of the operation of all the learned subregularities, so there might be some transfer from regular learning in this case. For example, the final Wickelphone in the correct past tense form of *hit*, namely *it#*, shares many of its Wickelfeatures with those of the regular past tense allomorphs such as *id#*. Let us call this effect *between-class transfer*.

It is important to note that much of the between-class transfer effect may be a consequence—perhaps even an artifact—of the Wickelfeature representation and one of the measures defined over it, namely percentage of incorrect Wickelfeatures activated in the output. Imagine that the model's learning component actually treated no-change verbs and other kinds of verbs identically, generating Wickelfeature sets of equal strength for *cutted* and *taked*. Necessarily, *taked* must contain more incorrect Wickelfeatures than *cutted*: most of the Wickelfeatures that one would regard as "incorrect" for *cutted*, such as those that correspond to the Wickelphone *tid* and *id#*, happen to characterize the *stem* perfectly (StopVowelStop, InterruptedFrontInterrupted, etc.), because *cut* and *ted* are featurally very similar. On the other hand, the incorrect Wickelfeatures for *taked* (those corresponding to Wickelphones *Akt* and *kt#*) will not characterize the correct output form *took*. This effect is exaggerated further by the fact that there are many more Wickelfeatures representing word boundaries than representing the same phonemes string-internally, as Lachter and Bever (1988) point out (recall that the Wickelfeature set was trimmed so as to exclude those whose two context phonemes belonged to different phonological dimensions—since the word-boundary feature # has no phonological properties, such a criterion will leave all Wickelfeatures of the form XY# intact). This difference is then carried over to the current implementation of the response-generation component, which puts response candidates at a disadvantage if they do not account for activated Wickelfeatures. The entire effect (a consequence of the fact that the model does not keep track of which features go in which positions) can be viewed either as a bug or a feature. On the one hand, it is one way of generating the (empirically correct) phenomenon that no-change responses are more common when stems have the same endings as the affixes that would be attached to them. On the other hand, it is part of a family of phonological confusions that result from the Wickelphone/Wickelfeature representations in general (see the section on Wickelphonology) and that hobble the model's ability even to reproduce strings verbatim. If the stem-affix feature confusions really are at the heart of the model's no-change responses, then it should also have recurring problems, unrelated to learning, in generating forms such as *pitted* or *pocketed* where the same Wickelfeatures occur in the stem and affix or even twice in the same stem but they must be kept distinct. Indeed, the model really does seems prone to make these undesirable errors, such as generating a single CVC sequence when two are necessary, as in the no-change responses for *hug*, *smoke*, and *brown*, or the converse, in overmarking errors such as *typeded* and *steppeded*.

A third possible reason that no-change responses are easy for *t*/*d*-final stems is that unlike other classes of irregulars in English, the no-change class has a single kind of change (that is, no change at all), and all its members have a phonological property in common: ending with a *t* or *d*. It is also the largest irregular subclass. The model has been given relatively consistent evidence of the contingency that verbs ending in *t* or *d* tend to have unchanged past tense forms, and it has encoded that contingency, presumably in large part by strengthening links between input Wickelfeatures representing word-final *t*/*d*s and identical corresponding output Wickelfeatures. Basically, the model is potentially sensitive to any statistical correlation between input and output feature sets, and it has picked up that one. That is, the acquisition of the simple contingency "end in *t*/*d* → no change" presumably makes the model mimic children. We can call this the *within-class uniformity* effect. As we have mentioned, the simplified rule-hypothesization mechanism presented in a previous section can acquire the same contingency (add a null affix for verbs ending in a nonsonorant noncontinuant coronal), and strengthen it with every no-change pair in the input. If, as we have argued, a rule-learning model considered many rules exemplified by input pairs before being able to determine which of them was the correct productive rule or rules for the language, this rule would exist in the child's grammar and would compete with the regular *d* rule and with other rules, just as competing outputs are computed in the RM model.

Finally, there is a fourth mechanism that was mentioned in our discussion of the strong verb system. Addition of the regular suffix *d* to a form ending in *t* or *d* produces a phonologically-illicit consonant cluster: *td* or *dd*. For regular verbs, the phonological rule of vowel insertion places an *i* between the two consonants. Interestingly, no irregular past ends in *id*, though some add a *t* or *d*. Thus we find *tell/told* and *leave/left*, but we fail to find *bleed/bledded* or *get/gotted*. A possible explanation is that a phonological rule, degemination, removes an affix after it is added as an alternative means of avoiding adjacent coronals in the strong class. The no-change verbs would then just be a special case of this generalization, where the vowel doesn't change either. Basically, the child would capitalize on a phonological rule acquired elsewhere in the system, and might overgeneralize by failing to restrict the degemination rule to the strong verbs.

Thus we have an overlapping set of explanations for the early acquisition and overgeneralization of the no-change contingency. Bybee and Slobin cite misperception, Rumelhart and McClelland cite between-class transfer and within-class uniformity, and rule-based theories can cite within-class uniformity or overgeneralized phonology. What is the evidence concerning the reasons that children are so sensitive to this contingency?

Unfortunately, a number of confounds in English make the theories difficult to distinguish. No-change verbs have a diagnostic phonological property in common with one another. They also share a phonological property with regular inflected past tense forms. Unfortunately, they are the same property: ending with *t/d*. And it is the sharing of that phonological property that triggers the putative phonological rule. So this massive confound prevents one from clearly distinguishing the accounts using the English past tense rule; one cannot say that the Rumelhart–McClelland model receives clear support from its ability to mimic children in this case.

In principle, a number of more diagnostic tests are possible. First, one must explain *why* the no-change class is confounded. The within-class uniformity account, which is one of the factors behind the RM model's success, cannot do this: if it were the key factor, we would surmise that English could just as easily have contained a no-change class defined by *any* easily-characterized within-class property (e.g. begin with *j*, end with *s*). Bybee and Slobin note that across languages, it is very common for no-change stems to contain the very ending that a rule would add. While ruling out within-class uniformity as the *only* explanation, this still leaves misperception, transfer, and phonology as possibilities, all of which foster learning of no-change forms for stems resembling the relevant affix.

Second, one can look at cases where possessing the features of the regular ending is not confounded with the characteristics of the no-change class. For example, the nouns that do not change when pluralized in English such as *sheep* and *cod* do not in general end in an *s* or *z* sound. If children nonetheless avoid pluralizing nouns like *ax* or *lens* or *sneeze*, it would support one or more of the accounts based on stem-affix similarity. Similarly, we might expect children to be reluctant to add *-ing* to form verbs like *ring* or *hamstring* or *rethink*.

If such effects were found, differences among verbs all of which resemble the affix in question could discriminate the various accounts that exploit the stem-affix similarity effect in different ways. Transfer, which is exploited by the RM model, would, all other things being equal, lead to equally likely no-change responses for all stems with a given degree of similarity to the affix. Phonology would predict that transfer would occur only when the result of adding an affix led to adjacent similar segments; thus it would predict more no-change responses for the plural of *ax* than the progressive of *sting*, which is phonologically acceptable without the intervention of any further rule.

Returning now to a possible unconfounded test of the within-class uniformity effect (implicated by Rumelhart and McClelland and by the rule-hypothesization module), one could look for some phonological property in common among a set of no-change stems that was independent of the phonological property of the relevant affix and see whether children were more likely to yield both correct and incorrect no-change responses when a stem had that property. As we have pointed out, monosyllabicity is a property holding of the irregular verbs in general, and of the no-change verbs in particular; presumably it is for this reason that the RM model, it turns out, is particularly susceptible to leaving regular verbs ending in *t/d* erroneously unchanged when they are monosyllabic. As Rumelhart and McClelland point out, if children are less likely to leave verbs such as *decide* or *devote* unchanged than verbs such as *cede* or *raid* it would constitute a test of this aspect of their theory; this test is not confounded by effects of across-class transfer.[29]

[29] Actually, this test is complicated by the fact that monosyllabicity and irregularity are not independent: in English, monosyllabicity is an important feature in defining the domain of many morphological and syntactic rules (e.g. *nicer/*intelligenter, give/*donate the museum a painting*; see Pinker, 1984), presumably because in English a monosyllable constitutes the minimal or basic word (McCarthy & Prince, forthcoming). As we have pointed out, all the irregular verbs in English are monosyllables or contain monosyllabic roots, (likewise for

A possible test of the misperception hypothesis is to look for other kinds of evidence that children misperceive certain stems as falling into a morphological category that is characteristically inflected. If so, then once the regular rule is acquired it could be applied in reverse to such misperceived forms, resulting in back-formations. For no-change verbs, this would result in errors such as *bea* or *blas* for *beat* or *blast*. We know of no reports of such errors among past tense forms (many would be impossible for phonological reasons) but have observed in Lisa's speech *mik* for *mix*, and in her noun system *clo (thes), len (s), sentent* (cf. *sentence*), *Santa Clau (s), upstair (s), downstair (s), bok* (cf. *box*), *trappy* (cf. *trapeze*), and *brefek* (cf. *brefeks* = 'breakfast').[30]

Finally, the process by which Rumelhart and McClelland exploit stem-affix similarity, namely transfer of the strength of the output features involved in regular pairs to the no-change stems, can be tested by looking at examples of blends of regular and subregular alternations that involve classes of verbs other than the no-change class. One must determine whether children produce such blends and whether it is a good thing or a bad thing for the RM theory that their model does so. We examine this issue in the next two sections.

In sum, the class of English verbs that do not change in the past tense involves a massive confound of within-class phonological uniformity and stem-affix similarity, leading to a complex nexus of predictions as to why children are so sensitive to the properties of the class. The relations between different models of past tense acquisition, predictions of which linguistic variables should have an effect on languages and on children, and the classes of verbs instantiating those variables, is many-to-many-to-many. Painstaking testing of the individual predictions using unconfounded sets of items in a variety of inflectional classes in English and other languages could tease the

nouns), a fact related in some way to irregularity being restricted to roots and monosyllables being prototypical English roots. So if children know that only roots can be irregular and that roots are monosyllables, (see Gordon, 1986, for evidence that children are sensitive to the interaction between roothood and morphology, and Gropen & Pinker, 1986, for evidence that they are sensitive to monosyllabicity), they may restrict their tendency to no-change responses to monosyllables even if it is not the product of their detecting the first-order correlation between monosyllabicity and unchanged pasts. Thus the ideal test would have to be done for some other language, in which a no-change class had a common phonological property independent of the definition of a basic root in the language, and independent of the phonology of the regular affix.

[30] Note that the facts of English do not comport well with any strong misperception account that would have the child invariably misanalyze irregular pasts as pseudo-stems followed by the regular affix: the majority of no-change verbs either have lax vowels and hence would leave phonologically impossible pseudo-stems after the affix was subtracted, such as *hi* or *cu*, or end in vowel-*t* sequences, which never occur in regular pasts and only rarely (e.g. *bought*) in irregulars. For the same reason it is crucial to Bybee and Slobin's account that children be constrained to form the schema ⟨*past: ...t/d#*⟩ rather than several schemas matching the input more accurately, such as ⟨*past: ...*[*unvoiced*] *t #*⟩] and ⟨*past: ...*[*voiced*] *d #*⟩. If they did, they would never misperceive *hit* and *cut* as past tense forms.

effects apart. At present, however, a full range of possibilities are all consistent with the data, ranging from the RM model explaining much of the phenomenon to its being entirely dispensable. The model's ability to duplicate children's performance, in and of itself, tells us relatively little.

7.2.3. Frequency of overregularizing irregular verbs in different vowel-change subclasses

Bybee and Slobin examined eight different classes of irregular past tense verbs (see the Appendix for an alternative, more fine-grained taxonomy). Their Class I contains the no-change verbs we have just discussed. Their Class II contains verbs that change a final *d* to *t* to form the past tense, such as *send/sent* and *build/built*. The other six classes involve vowel changes, and are defined by Bybee and Slobin as follows:

- Class 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.*
- Class 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.* [Bybee and Slobin include in this class the pair *buy/bought* even though it does not involve a deleted consonant. *Make/made* and *have/had* were also included even though they do not involve a vowel change.]
- Class V. Verbs that undergo an internal vowel change whose stems end in a dental, e.g. *bite/bit, find/found, ride/rode.*
- Class VI. Verbs that undergo a vowel change of /I/ to /æ/ or /ʌ/, e.g. *sing/sang, sting/stung.*
- Class VII. All other verbs that undergo an internal vowel change, e.g. *give/gave, break/broke.*
- Class VIII. All verbs that undergo a vowel change and that end in a diphthongal sequence, e.g. *blow/blew, fly/flew.* [*Go/went* is also included in this class.]

Bybee and Slobin noted that preschoolers had widely varying tendencies to overregularize the verbs in these different classes, ranging from 10% to 80% of the time (see the first column of Table 2). Class IV and III verbs, whose past tense forms receive a final *t/d* in addition to their vowel changes, were overregularized the least; Class VII and V verbs, which have unchanged final consonants and a vowel change, were overregularized somewhat more often; Class VI verbs, involving the *ing-ang-ung* regularity, were regularized more often than that; and Class VIII verbs, which end in a diphthong sequence which is changed in the past, were overregularized most often. Bybee and Slobin again account for this phenomenon by appealing to factors affecting the process of juxtaposing corresponding present and past forms. They suggest that the presence of an added *t/d* facilitates the child's recognition that Class III and IV past forms *are* past forms, and that the small percentage of shared segments between Class VIII present and past versions (e.g., one for *see/saw* or *know/knew*) hinders that recognition process. As the likelihood of successful juxtaposition of present and past forms decreases, the likelihood of the regular rule to operate, unblocked by an irregular past form, increases and overregularizations become more common.

Rumelhart and McClelland suggest, as in their discussion of no-change verbs, that their model as it stands can reproduce the developmental phenomenon. Since the Bybee and Slobin subjects range from $1\frac{1}{2}$ to 5 years, it is not clear which stage of performance of the model should be compared with that of the children, so Rumelhart and McClelland examined the output of the model at several stages. These stages corresponded to the model's first five trials with the set of medium-frequency, predominantly regular verbs, the next five trials, the next ten trials, and an average over those first twenty trials (these intervals constitute the period in which the tendency of the model to overregularize was highest). The average strength of the overregularized forms within each class was calculated for each of these four intervals.

The fit between model and data is good for the interval comprising the first five trials, which Rumelhart and McClelland concentrate on. We calculate the rank-order correlation between degree of overregularization by children and model across classes as .77 in that first interval; however it then declines to .31 and .14 in the next two intervals and is .31 for the average response over all three intervals. The fact that the model is only successful at accounting for Bybee and Slobin's data for one brief interval (less than 3% of the training run) selected post hoc, whereas the data themselves are an average over a span of development of $3\frac{1}{2}$ years, should be kept in mind in evaluating the degree of empirical confirmation this study gives the model. Nonetheless, the tendency of Class VIII verbs (*fly/flew*) to be most often regularized, and for Class III verbs (*feel/felt*) to be among those least often regularized, persists across all three intervals.

The model, of course is insensitive to any factor uniquely affecting the juxtaposition of present and past forms because such juxtaposition is accomplished by the "teacher" in the simulation run. Instead, its fidelity to children's overregularization patterns at the very beginning of its own overregularization stage must be attributed to some other factor. Rumelhart and McClelland point to differences among the classes in the frequency with which their characteristic vowel changes are exemplified by the verb corpus as a whole. Class VIII verbs have vowel shifts that are relatively idiosyncratic to the individual verbs in the class; the vowel shifts of other classes, on the other hand, might be exemplified by many verbs in many classes. Furthermore, Class III and IV verbs, which require the addition of a final *t/d*, can benefit from the fact that the connections in the network that effect the addition of a final *t/d* have been strengthened by the large number of regular verbs. The model creates past tense forms piecemeal, by links between stem and past Wickelfeatures, and with no record of the structure of the individual words that contributed to the strengths of those links. Thus vowel shifts and consonant shifts that have been exemplified by large numbers of verbs can be applied to different parts of a base form even if the exact combination of such shifts exemplified by that base form is not especially frequent.

How well could the simplified rule-finding module account for the data? Like the RM model, it would record various subregular rules as candidates for a regular past tense rule. Assuming it is sensitive to type frequency, the rule candidates for more-frequently exemplified subregularities would be stronger. And the stronger an applicable subregular rule candidate is, the less is the tendency for its output to lose the competition with the overregularized form contributed by the regular rule. Thus if Rumelhart and McClelland's explanation of their model's fit to the data is correct, a rule-finding model sensitive to type-frequency presumably would fit the data as well.

This conjecture is hard to test because Bybee and Slobin's data are tabulated in some inconvenient ways. Each class is heterogeneous, containing verbs governed by a variety of vowel-shifts and varying widely as to the number of such shifts in the class and the number of verbs exemplifying them within the class and across the classes. Furthermore, there are some quirks in the classification. *Go/went*, the most irregular main verb in English, is assigned to Class VIII, which by itself could contribute to the poor performance of children and the RM model on that class. Conversely, *have* and *make*, which involve no vowel shift at all, are included in Class IV, possibly contributing to good average performance for the class by children and the model. (See the Appendix for an alternative classification.)

It would be helpful to get an estimate as to how much of the RM model's empirical success here might be due to the different frequencies of exemplification of the vowel-shift subregularities within each class, because such an effect carries over to a symbolic rule-finding alternative. To get such an estimate, we considered each vowel shift(e.g. *i* → *æ*) as a separate candidate rule, strengthened by a unit amount with each presentation of a verb that exemplifies it in the Rumelhart–McClelland corpus of high- and medium-frequency verbs. To allow *have* and *make* to benefit from the prevalence of other verbs whose vowels do not change, we pooled the different vowel no-change rules (*a* → *a; i* → *i*, etc.) into a single rule (the RM model gets a similar benefit by using Wickelfeatures, which can code for the presence of vowels, rather than Wickelphones) whose strength was determined by the number of no-vowel-change verbs in Classes I and II.[31] Then we averaged the strengths of all the subregular rules included within each of Bybee and Slobin's classes. These averages allow a prediction of the ordering of overregularization probabilities for the different subclasses, based solely on the number of irregular verbs in the corpus exemplifying the specific vowel alternations among the verbs in the class. Though the method of prediction is crude, it is just about as good at predicting the data as the output of the RM model during the interval at which it did best and much better than the RM model during the other intervals examined. Specifically, the rank-order correlation between number of verbs in the corpus exemplifying the vowel shifts in a class and the frequency of children's regularization of verbs in the class is .71. The data, predictions of the RM model, and predictions from our simple tabulations are summarized in Table 2.

What about the effect of the addition of *t/d* on the good performance on Class III and IV verbs? The situation is similar in some ways to that of the no-change verbs discussed in the previous section. The Class III and IV verbs

Table 2. *Ranks of tendencies to overregularize irregular verbs involving vowel shifts*

Verb subclass		Children*	RM 1st set	RM 2nd set	RM 3rd set	RM average	Avg. freq. of vowel shift**
VIII	blow/blew	1 (.80)	1	1	1	1	1 (1.6)
VI	sing/sang	2 (.55)	4	4	4	4	3 (2.7)
V	bite/bit	3 (.34)	2	3	6	3	5 (3.9)
VII	break/broke	4 (.32)	3	6	3	6	2 (2.1)
III	feel/felt	5 (.13)	6	5	5	5	4 (3.8)
IV	seek/sought	6 (.10)	5	2	2	2	6 (4.5)
Rank order correlation with children's proportions			.77	.31	.14	.31	.71

* Actual proportions of regularizations by children are in parentheses.
** Mean number of verbs in the irregular corpus exemplifying the vowel shifts within a class are indicated in parentheses.

[31] In a sense, it would have been more accurate to calculate the strength of the no-vowel-change rule on the basis of all the verbs in the corpus, regular and irregular, rather than just the irregular verbs. But with our overly simple strength function, this would have greatly stacked the deck in favor of correctly predicting low regularization rates for Class IV verbs and so we only counted the exemplifications of no-vowel-change within the irregular verbs.

take some of the most frequently-exemplified vowel-changes (including no-change for *have* and *make*); they also involve the addition of *t* or *d* at the end causing them to resemble the past tense forms of regular verbs. Given this confound, good performance with these classes can be attributed to either factor and so the RM model's good performance with them does not favor it over the Bybee-Slobin account focusing on the juxtaposition problem.

The question of blended responses. An interesting issue arises, however, when we consider the possible effects of the addition of *t/d in combination with* the effects of a common vowel shift. Recall that the RM model generates its output piecemeal. Thus strong regularities pertaining to different parts of a word can affect the word simultaneously, producing a chimerical output that need not correspond in its entirety to previous frequent patterns. To take a simplified example, after the model encounters pairs such as *meet/met* it has strong links between *i* and ɛ; after it encounters pairs such as *play/played* it has strong links between final vowels and final vowel-*d* sequences; when presented with *flee* it could then generate *fled* by combining the two regularities, even if it never encountered an *ee/ed* alternation before. What is interesting is that this blending phenomenon is the direct result of the RM model's lack of word structure. In an alternative rule-finding account, there would be an *i* → ɛ rule candidate, and there would be a *d*-affixation rule candidate, but they would generate two distinct competing outputs, not a single blended output. (It is possible in principle that some of the subregular strong verbs such as *told* and *sent* involve the superposition of independent subregular rules, especially in the history of the language, but in modern English one cannot simply heap the effect of the regular rule on top of any subregular alternation, as the RM model is prone to do.) Thus it is not really fair for us to claim that a rule-hypothesization model can account for good performance with Class III and IV verbs because they involve frequently-exemplified vowel alternations; such alternations only result in correct outputs if they are blended with the addition of a *t/d* to the end of the word. In principle, this could give us a critical test between the network model and a rule-hypothesization model: unlike the ability to soak up frequent alternations, the automatic superposition of any set of them into a single output is (under the simplest assumptions) unique to the network model.

This leads to two questions: Is there independent evidence that children blend subregularities? And does the RM model itself really blend subregularities? We will defer answering the first question until the next section, where it arises again. As for the second, it might seem that the question of whether response blending occurs is perfectly straightforward, but in fact it is not. Say the model's active output Wickelfeatures in response to *flee* include those for medial ε and those for word-final *d*. Is the overt response of the model *fled*, a correct blend, or does it set up a competition between [*flid*] and [*flε*], choosing one of them, as the rule-hypothesization model would? In principle, either outcome is possible, but we are never given the opportunity to find out. Rumelhart and McClelland do not test their model against the Bybee and Slobin data by letting it output its favored response. Rather, they externally assemble alternatives corresponding to the overregularized and correct forms, and assess the relative strengths of those alternatives by observing the outcome of the competition in the restricted-choice whole-string binding network (recall that the output of the associative network, a set of activated Wickelfeatures, is the input to the whole-string binding network). These strengths are determined by the number of activated Wickelfeatures that each is consistent with. The result is that correct alternatives that also happen to resemble blends of independent subregularities are often the response chosen. But we do not know whether the model, if left to its own devices, would produce a blend as its top-ranked response.

Rumelhart and McClelland did not perform this test because it would have been too computationally intensive given the available hardware: recall that the only way to get the model to produce a complete response form on its own is by giving it (roughly) all possible output strings (that is, all permutations of segments) and having them compete against each other for active Wickelfeatures in an enormous "unconstrained whole-string binding network". This is an admitted kluge designed to give approximate predictions of the strengths of responses that a more realistic output mechanism would construct. Rumelhart and McClelland only ran the unconstrained whole-string binding network on a small set of new low-frequency verbs in a transfer test involving no further learning. It is hard to predict what will happen when this network operates because it involves a "rich-get-richer" scheme in the competition among whole strings, by which a string that can uniquely account for some Wickelfeatures (including Wickelfeatures incorrectly turned on as part of the noisy output function) gets disproportionate credit for the features that it and its competitors account for equally well, occasionally leading to unpredictable winners. In fact, the whole-string mechanism does yield blends such as *slip/slept*. But as mentioned, these blends are also occasionally bizarre, such as *mailed/membled* or *tour/toureder*. And this is why the question of overt blended outputs is foggy: it is unclear whether tuning the whole-string binding network, or a more reasonable output construction mechanism, so that the bizarre blends were eliminated, would also eliminate the blends that perhaps turn out to be the correct outputs for Class III and IV.[32]

[32] To complicate matters even further, even outright blends are possible in principle within the rule-based

In sum, the relative tendencies of children to overregularize different classes of vowel-change irregular verbs does not favor the RM model. The model for one brief stage selected post hoc shows a moderately high correlation with data on children's behavior in the strength it assigns to overregularized forms. Much of this correlation is simply due to the frequencies with which the vowel alternations in a given class have been exemplified by verbs in the corpus as a whole. A rule-hypothesization model would also be sensitive to these frequencies under even the simplest of assumptions. But the ability of the network model to blend independent subregularities into a single response follows naturally from its lack of word structure and could lead to tests distinguishing the models. Unfortunately, whether the network model would actually output blended responses in its best incarnation is unknown; whether children output blended responses is a question we will turn to shortly.

7.2.4. "Eated" versus "ated" errors

The final developmental phenomenon that Rumelhart and McClelland examine is the tendency of children to produce overregularization errors consisting of an irregular past affixed with *ed*, such as *ated* or *broked*. Such errors tend to occur considerably later in development than errors consisting of a base form affixed with *ed*, such as *eated* or *breaked* (Kuczaj, 1977). Rumelhart and McClelland compared the strength of *eated* and *ated* outputs for the irregular verbs in their corpus. They found that the strength of the *ated* form relative to the *eated* form increased over the course of training, thus mimicking the Kuczaj data.

What causes past + *ed* errors? There are two possibilities. One is that the child sometimes fails to realize that the irregular past is the past tense form of some base form. Thinking it is a base form itself, he or she feeds it into the past tense formation mechanism and gets a doubly-marked error. This cannot be the explanation for the model's behavior because correct present/past pairs are always provided to it. The alternative is that the two different changes are applied to the correct base form and the results are blended to yield the double-marking. This is similar to one of the explanations for the model's relatively infrequent overregularization of Class III and IV verbs discussed in the previous section, and to one of the explanations for the

model. For example, children might have two subregular rules that both apply, as might have been appropriate in an earlier stage of English. Or, there may be a response buffer that receives the output of the competition process, and occasionally two candidates of approximately equal strength slip out of the competition mechanism and are blended in the response buffer. The result would be a blended speech error from the point of view of the "design" of the rule-application module but possibly an adventitious correct response. Though this account may not seem as natural as the blending inherent in the network model, the notion of a serially ordered response buffer distinct from a representation of target segments is part of standard explanations for anticipatory and perseverative speech errors (e.g., Shattuck-Hufnagel, 1979).

model's tendency to leave *t/d*-final stems unchanged discussed in the section before that. As in the previous discussions, the lack of a realistic response production mechanism makes it unclear whether the model would ever actually produce past + *ed* blends when it is forced to utter a response on its own, or whether the phenomenon is confined to such forms simply increasing in strength in the three-alternative forced-choice experiment because only the past + *ed* form by definition contains three sets of features all of them strengthened in the course of learning (its idiosyncratic features, the features output by subregularities, and the features of regularized forms). In Rumelhart and McClelland's transfer test on new verbs, they chose a minimum strength value of .2 as a criterion for when a form should be considered as being a likely overt response of the model. By this criterion, the model should be seen as rarely outputting past + *ed* forms, since such forms on the average never exceed a strength of .15. But let us assume for now that such forms would be output, and that blending is their source.

At first one might think that the model had an advantage in that it is consistent with the fact that *ated* errors increase relative to the *eated* errors in later stages, a phenomenon not obviously predicted by the misconstrued stem account.[33] However, many of the phenomena we discuss below that favor the misconstrued-stem account over the RM model appear during the same, relatively late period as the *ated* errors (Kuczaj, 1981), so lateness itself does not distinguish the accounts. Moreover, Kuczaj (1977) warns that the late-*ated* effect is not very robust and is subject to individual differences. In a later study (Kuczaj, 1978), he eliminated these sampling problems by using an experimental task in which children judged whether various versions of past tense forms sounded "silly". He found in two separate experiments that while children's acceptance of the *eated* forms declined monotonically their acceptance of *ated* forms showed an inverted U-shaped function, first increasing *but then decreasing* relative to *eated*-type errors. Since in the RM model the strengths of both forms monotonically approach an asymptote near zero, with the curves crossing only once, the model demonstrates no special ability to track the temporal dynamics of the two kinds of errors. In the discussion below we will concentrate on the reasons that such errors occur in the first place.

Once again, a confound in the materials provided by the English language confounds Rumelhart and McClelland's conclusion that their model accounts

[33] Kuczaj did suggest an interesting hypothesis: children might treat *went* as a base form that expresses pastness inherently, as part of its intrinsic meaning, rather than as the grammatical composition of *go* + *past* (Kuczaj, 1981, observed *was wenting* but never *is wenting*). The eventual realization that tense must be marked *grammatically*, and not just implicitly by the inherent meanings of verbs, is a later acquisition, and its effect is the regular inflection of irregular forms.

children's *ated*-type errors. Irregular past tense forms appear in the child's input and hence can be misconstrued as base forms. They also are part of the model's output for irregular base forms and hence can be blended with the regularized response. Until forms which have one of these properties and not the other are examined, the two accounts are at a stalemate.

Fortunately, the two properties can be unconfounded. Though the correct irregular past will usually be the strongest non-regularized response of the network, it is also sensitive to subregularities among vowel changes and hence one might expect blends consisting of a frequent and consistent but incorrect vowel change plus the regular *ed* ending. In fact the model does produce such errors for regular verbs it has not been trained on, such as *shape/shipped, sip/sepped, slip/slept*, and *brown/brawned*. Since the stems of these responses are either not English verbs or have no semantic relationship to the correct verb, such responses can never be the result of mistakenly feeding the wrong base form of the verb into the past tense formation process. Thus if the blending assumed in the Rumelhart and McClelland model is the correct explanation for children's past + *ed* overregularizations, we should see children making these and other kinds of blend errors. We might also expect errors consisting of a blend of a correct irregular alteration of a verb plus a frequent subregular alteration, such as *send/soant* (a blend of the *d* → *t* and ε → *o* subregularities) or *think/that* (a blend of the *ing* → *ang* and *final consonant cluster* → *t* subregularities). (As mentioned, though, these last errors are not ruled out in principle in all rule-based models, since superposition may have had a role in the creation of several of the strong past forms in the history of English, but indiscriminately adding the regular affix onto strong pasts is ruled out by most theories of morphology.)

Conversely, if the phenomenon is due to incorrect base input forms, we might expect to see other inflection processes applied to the irregular past, resulting in errors such as *wenting* and *broking* or *wents* and *brokes*. Since mechanisms for progressive or present indicative inflection would never be exposed to the idiosyncrasies or subregularities of irregular past tense forms under Rumelhart and McClelland's assumptions, such errors could not result from blending of outputs. Similarly, irregular pasts should appear in syntactic contexts calling for bare stems if children misconstrue irregular pasts as stems. In addition, we might expect to find cases where *ed* is added to incorrect base forms that are plausible confusions of the correct base form but implausible results of the mixing of subregularities.

Finally, we might expect that if children are put in a situation in which the correct stem of a verb is provided for them, they would not generate past + *ed* errors, since the source of such errors would be eliminated.

All five predictions work against the RM model and in favor of the explanation based on incorrect inputs. Kuczaj (1977) reports that his transcripts contained no examples where the child overapplied any subregularity, let alone a blend of two of them or of a subregularity plus the regular ending. Bybee and Slobin do not report any such errors in children's speech, though they do report them as adult slips of the tongue in a time-pressured speaking task designed to elicit errors. We examined the full set of transcripts of Adam, Eve, Sarah, and Lisa for words ending in *-ed*. We found 13 examples of irregular past + *ed* or *en* errors in past and passive constructions:

(36) Adam: ranned
tooked
stoled (twice)
broked (participle)
felled

Eve: tored

Sarah: flewed (twice)
caughted
stucked (participle)

Lisa: torned (participle)
tooken (twice) (participle)
sawn (participle)

The participle forms must be interpreted with caution. Because English irregular participles sometimes consist of the stem plus *en* (e.g. *take - took - taken*) but sometimes consist of the irregular past plus *en* (e.g. *break - broke - broken*), errors like *tooken* could reflect the child overextending this regularity to past forms of verbs that actually follow the stem + *en* pattern; the actual stem or even the child's mistaken hypothesis about it may play no role.

What about errors consisting of a subregular vowel alternation plus the addition of *ed*? The only examples where an incorrect vowel other than that of the irregular form appeared with *ed* are the following:

(37) Adam: I think it's not fulled up to de top.
I think my pockets gonna be all fulled up.
I'm gonna ask Mommy if she has any more grain ... more stuff that she needs grained. [He has been grinding crackers in a meat grinder producing what he calls "grain".]

Sarah: Oo, he hahted.

Lisa: I brekked your work.

For Adam, neither vowel alternation is exemplified by any of the irregular verbs in the Rumelhart-McClelland corpus, but in both cases the stem is identical to a non-verb that is phonologically and semantically related to the target verb and hence may have been misconstrued as the base form of the verb or converted into a new base verb. Sarah's error can be attributed directly to phonological factors since she also pronounced *dirt*, involving no morphological change, as "dawt", according to the transcriber. This leaves Lisa's *brekked* as the only putative example; note that unlike her single-subregularity errors such as *bote* for *bit* which lasted for extended periods of time, this appeared only once, and the correct form *broke* was very common in her speech. Furthermore blending is not a likely explanation: among high and middle frequency verbs, the alternation is found only in *say/said* and to a lesser extent in pairs such as *sleep/slept* and *leave/left*, whereas in many other alternations the *e* sound is mapped onto other vowels (*bear/bore, wear/wore, tear/tore, take/took*, and *shake/shook*). Thus it seems unlikely that the RM model would produce a blend in this case but not in the countless other opportunities for blending that the children avoided. Finally, we note that Lisa was referring to a pile of papers that she scattered, an unlikely example of *breaking* but a better one of *wrecking*, which may have been the target serving as the real source of the blend (and not a past tense subregularity) if it was a blend. In sum, except perhaps for this last example under an extremely charitable interpretation, the apparent blends seem far more suggestive of an incorrect stem correctly inflected than a blend between two past tense subregularities.

This conclusion is strengthened when we note that children do make errors such as *wents* and *wenting*, which could only result from inflecting the wrong stem. Kuczaj (1981) reports frequent use of *wenting, ating*, and *thoughting* in the speech of his son, and we find in Eve's speech *fells* and *wents* and in Lisa's speech *blow awayn, lefting, hidding* (= hiding), *stoling, to took, to shot*, and *might loss*. These last three errors are examples of a common phenomenon sometimes called 'overtensing', which because it occurs mostly with irregulars (Maratsos & Kuczaj, 1978), is evidence that irregulars are misconstrued as stems (identical to infinitives in English). Some examples from Pinker (1984) include *Can you broke those, What are you did?, She gonna fell out*, and *I'm going to sit on him and made him broken*. Note that since many of these forms occur at the same time as the *ated* errors, the relatively late appearance of *ated* forms may reflect the point at which stem extraction (and mis-extraction) in general is accomplished.

Finally, Kuczaj (1978) presents more direct evidence that past + *ed* errors are due to irregular pasts misconstrued as stems. In one of his tasks, he had children convert a future tense form (i.e. "X will + ⟨verb stem⟩") into a past tense form (i.e. "X already ⟨verb past⟩"). Past + *ed* errors virtually vanished (in fact they completely vanished for two of the three age groups). Kuczaj argues that the crucial factor was that children were actually given the proper base forms. This shows that children's derivation of the errors must be from *ate* to *ated*, not, as it is in the RM model, from *eat* to *ated*.

Yet another test of the source of apparently blended errors is possible when we turn our attention to the regular system. If the child occasionally misanalyzes a past form as a stem, he or she should do so for regular inflected past forms and not just irregular ones, resulting in errors such as *talkeded*. The RM model also produces such errors as blends, but for reasons that Rumelhart and McClelland do not explain, all these errors involve regular verbs whose stems end in *p* or *k*: *carpeded, drippeded, mappeded, smokeded, snappeded, steppeded*, and *typeded*, but not *browneded, warmeded, teareded* or *clingeded*, nor, for that matter, irregular stems of any sort: the model did not output *creepeded/crepted, weepeded/wepted, diggeded*, or *stickeded*. We suggest the following explanation for this aspect of the model's behavior. The phonemes *p* and *k* share most of their features with *t*. Therefore on a Wickelfeature by Wickelfeature basis, learning that *t* and *d* give you *id* in the output transfers to *p, b, g* and *k* as well. So there will be a bias toward *id* responses after all stops. Since there is also a strong bias toward simply adding *t*, there will be a tendency to blend the 'add *t*' and the 'add *id*' responses. Irregular verbs, as we have noted, never end in *id*, so to the extent that the novel irregulars resemble trained ones (see Section 4.4), the features of the novel irregulars will inhibit the response of the *id* Wickelfeatures and double-marking will be less common.

In any case, though Rumelhart and McClelland cannot explain their model's behavior in this case, they are willing to predict that children as well will double-mark more often for *p*- and *k*-final stems. In the absence of an explanation as to why the model behaved as it did, Rumelhart and McClelland should just as readily extrapolate the model's reluctance to double-mark irregular stems and test the prediction that children should double-mark only regular forms (if our hypothesis about the model's operation is correct, the two predictions stem from a common effect). Checking the transcripts, we did find *ropeded* and *stoppeded* (the latter uncertain in transcription) in Adam's speech, and *likeded* and *pickeded* in Sarah's, as Rumelhart and McClelland would predict. But Adam also said *tieded* and Sarah said *buyded* and *makeded* (an irregular). Thus the model's prediction that double-marking should be specific to stems ending with *p* and *d*, and then only when they are regular, is not borne out. In particular, note that *buyded* and *tieded* cannot be the result of a blend of subregularities, because there *is* no subregularity according to which *buy* or *tie* would tend to attract a *id* ending.[34]

Finally, Slobin (1985) notes that Hebrew contains two quite pervasive rules for inflecting the present tense, the first involving a vowel change, the second a consonantal prefix and a different vowel change. Though Israeli children overextend the prefix to certain verbs belonging to the first class, they never blend this prefix of the second class with the vowel change of the first class. This may be part of a larger pattern that children seem to respect the integrity of the word as a cohesive unit, one that can have affixes added to it and that can be modified by general phonological processes, but that cannot simply be composed as a blend of bits and pieces contributed by various regular and irregular inflectional regularities. It is suggestive in this regard that Slobin (1985), in his crosslinguistic survey, lists examples from the speech of children learning Spanish, French, German, Hebrew, Russian, and Polish, where the language mandates a stem modification plus the addition of an affix and children err by only adding the affix.

Once again we see that the model does not receive empirical support from its ability to mimic a pattern of developmental data. The materials that Rumelhart and McClelland looked at are again confounded in a way that leaves their explanation and the standard one focusing on the juxtaposition problem equally plausible given only the fact of *ated* errors. One can do better than that. By looking at unconfounded cases, contrasting predictions leading to critical tests are possible. In this case, six different empirical tests all go against the explanation inherent in the Rumelhart and McClelland model: absence of errors due to blending of subregularities, presence of *wenting*-type errors, presence of errors where irregular pasts are used in nonpast contexts, presence of errors where the regular past ending is mistakenly applied to non-verb stems, drastic reduction of *ated*-errors when the correct stem is supplied to the child, and presence of errors where the regular ending is applied twice to stems that are irregular or that end in a vowel. These tests show that errors such as *ated* are the result of the child incorrectly feeding *ate* as a base form into the past tense inflection mechanism, and not the result of blending components of *ate* and *eated* outputs.

[34] One might argue that the misconstrued-stem account would fail to generate these errors, too, since it would require that the child first generate *maked* and *buyed* using a productive past tense rule and then forget that the forms really were in the past tense. Perhaps, the argument would go, some other kind of blending caused the errors, such as a mixture of the two endings *d* and *id* which are common across the language even if the latter is contraindicated for these particular stems. In fact, the misconstrued-stem account survives unscathed, because one can find errors not involving overinflection where child-generated forms are treated as stems: for example, Kuczaj (1976) reports sentences such as *They wouldn't haved a house* and *She didn't goed*.

7.3. Summary of how well the model fares against the facts of children's development

What general conclusions can we make from our examination of the facts of children's acquisition of the English past tense form and the ability of the RM model to account for them? This comparison has brought several issues to light.

To begin with, one must reject the premise that is implicit in Rumelhart and McClelland's arguments, namely that if their model can duplicate a phenomenon, the traditional explanation of that phenomenon can be rejected. For one thing, there is no magic in the RM model duplicating correlations in language systems: the model can extract any combination of over 200,000 atomic regularities, and many regularities that are in fact the consequences of an interaction among principles in several grammatical components will be detectable by the model as first-order correlations because they fall into that huge set. As we argued in Section 4, this leaves the structure and constraints on the phenomena unexplained. But in addition, it leaves many of the simple goodness-of-fit tests critically confounded. When the requirements of a learning system designed to attain the adult state are examined, and when unconfounded tests are sought, the picture changes.

First, some of the developmental phenomena can be accounted for by any mechanism that keeps records of regularities at several levels of generality, assigns strengths to them based on type-frequency of exemplification, and lets them compete in producing past tense candidate forms. These phenomena include children's shifts or waffling between irregular and overregularized past tense forms, their tendency not to change verbs ending in *t/d*, and their tendency to overregularize verbs with some kinds of vowel alternations less than others. Since there are good reasons why rule-hypothesization models should be built in this way, these phenomena do not support the RM model as a whole or in contrast with rule-based models in general, though they do support the more general (and uncontroversial) assumption of competition among multiple regularities of graded strength during acquisition.

Second, the lack of structures corresponding to distinct words in the model, one of its characteristic features in contrast with rule-based models, might be related to the phenomenon of blended outputs incorporating independent subregularities. However, there is no good evidence that children's correct responses are ever the products of such blends, and there *is* extensive evidence from a variety of sources that their *ated*-type errors are *not* the products of such blends. Furthermore, given that many blends are undesirable, it is not clear that the model should be allowed to output them when a realistic model of its output process is constructed.

Third, the three-stage or U-shaped course of development for regular and irregular past tense forms in no way supports the RM model. In fact, the model provides the wrong explanation for it, making predictions about changes in the mixture of irregular and regular forms in children's vocabularies that are completely off the mark.

This means that in the two hypotheses for which unconfounded tests are available (the cause of the U-shaped overregularization curve, and the genesis of *ated*-errors), both of the processes needed by the RM model to account for developmental phenomena—frequency-sensitivity and blending—have been shown to play no important role, and in each case, processes appealing to rules—to the child's initial hypothesization of a rule in one case, and to the child's misapplication of it to incorrect inputs in a second—have received independent support. And since the model's explanations in the two confounded cases (performance with no-change verbs, and order of acquisition of subclasses) appeal in part to the blending process, the evidence against blending in our discussion of the *ated* errors taints these accounts as well. We conclude that the developmental facts discussed in this section and the linguistic facts discussed in Section 4 converge on the conclusion that knowledge of language involves the acquisition and use of symbolic rules.

8. General discussion

Why subject the RM model to such painstaking analysis? Surely few models of any kind could withstand such scrutiny. We did it for two reasons. First, the conclusions drawn by Rumelhart and McClelland—that PDP networks provide exact accounts of psychological mechanisms that are superior to the approximate descriptions couched in linguistic rules; that there is no induction problem in their network model; that the results of their investigation warrant revising the way in which language is studied—are bold and revolutionary. Second, because the model is so explicit and its domain so rich in data, we have an unusual opportunity to evaluate the Parallel Distributed Processing approach to cognition in terms of its concrete technical properties rather than bland generalities or recycled statements of hopes or prejudices.

In this concluding section we do four things: we briefly evaluate Rumelhart and McClelland's strong claims about language; we evaluate the general claims about the differences between connectionist and symbolic theories of cognition that the RM model has been taken to illustrate; we examine some of the ways that the problems of the RM model are inherently due to its PDP architecture, and hence ways in which our criticisms implicitly extend to certain kinds of PDP models in general; and we consider whether the model could be salvaged by using more sophisticated connectionist mechanisms.

8.1. On Rumelhart and McClelland's strong claims about language

One thing should be clear. Rumelhart and McClelland's PDP model does not differ from a rule-based theory in providing a more exact account of the facts of language and language behavior. The situation is exactly the reverse. As far as the adult steady state is concerned, the network model gives a crude, inaccurate, and unrevealing description of the very facts that standard linguistic theories are designed to explain, many of them in classic textbook cases. As far as children's development is concerned, the model's accounts are at their best no better than those of a rule-based theory with an equally explicit learning component, and for two of the four relevant developmental phenomena, critical empirical tests designed to distinguish the theories work directly against the RM model's accounts but are perfectly consistent with the notion that children create and apply rules. Given these empirical failings, the ontological issue of whether the PDP and rule-based accounts are realisi portrayals of actual mechanisms as opposed to convenient approximate summaries of higher-order regularities in behavior is rather moot.

There is also no basis for Rumelhart and McClelland's claim that in their network model, as opposed to traditional accounts, "there is no induction problem". The induction problem in language acquisition consists, among other things, of finding sets of inputs that embody generalizations, extracting the right kinds of generalizations from them, and deciding which generalizations can be extended to new cases. The model does not deal at all with the first problem, which involves recognizing that a given word encodes the past tense and that it constitutes the past tense version of another word. This juxtaposition problem is relegated to the model's environment (its "teacher"), or more realistically, some unspecified prior process; such a division of labor would be unproblematic if it were not for the fact that many of the developmental phenomena that Rumelhart and McClelland marshall in support of their model may be intertwined with the juxtaposition process (the onset of overregularization, and the source of *ated* errors, most notably). The second part of the induction problem is dealt with in the theory the old-fashioned way: by providing it with an innate feature space that is supposed to be appropriate for the regularities in that domain. In this case, it is the distinctive features of familiar phonological theories, which are incorporated into the model's Wickelfeature representations (see also Lachter & Bever, 1987). Aspects in which the RM model differs from traditional accounts in how it uses distinctive features, such as representing words as unordered pools of feature trigrams, do not clearly work to the advantage of the model, to put it mildly. Finally, the theory deals very poorly with the crucial third aspect of the induction problem, when to generalize to new items. It cannot make proper phonological generalizations or respect the morphosyntactic constraints on the domain of application of the regular rule, and in its actual performance it errs in two ways, both overestimating the significance of the irregular subgeneralizations and underestimating the generality of the regular rule.

The third claim, that the success of their model calls for a revised understanding of language and language acquisition, is hardly warranted in light of the problems we have discussed. To give credit where it is due, we do not wish to deny the extent to which Rumelhart and McClelland's work has increased our understanding of language acquisition. The model has raised intriguing questions about the role of the family resemblance structure of subregularities and of their frequency of exemplification in overregularization, the blending of independent subregularities in generating overt outputs, effects of the mixture of regular and irregular forms in the input on the tradeoffs between rote and generalization, and the causes of transitions between developmental stages, in particular, the relative roles of the present-past juxtaposition process and the pattern-extraction process. But the model does not give superior or radically new answers for the questions it raises.

8.2. Implications for the metatheory and methodology of connectionism

Often the RM model is presented as a paradigm case not only of a new way to study language, but of a new way to understand what a cognitive theory is a theory of. In particular, a persistent theme in connectionist metatheory affirms that 'macro-level' symbolic theories can at best provide an approximate description of the domain of inquiry; they may be convenient in some circumstances, the claim goes, but never exact or real:

> Subsymbolic models accurately describe the microstructure of cognition, while symbolic models provide an approximate description of the macrostructure. (Smolensky, in press, p. 21)

> We view macrotheories as approximations to the underlying microstructure which the distributed model presented in our paper attempts to capture. As approximations they are often useful, but in some situations it will turn out that an examination of the microstructure may bring much deeper insight. (Rumelhart & McClelland, PDPI, p. 125)

> ... these [macro-level] models are approximations and should not be pushed too far. (Rumelhart & McClelland, p. 126; bracketed material ours here and elsewhere)

In such discussions the relationship between Newtonian physics and Quantum Mechanics typically surfaces as the desired analogy.

One of the reasons that connectionist theorists tend to reserve no role for higher-level theories as anything but approximations is that they create a dichotomy that, we think, is misleading. They associate the systematic, rule-based analysis of linguistic knowledge with what they call the "explicit inaccessible rule" view of psychology, which

> ... 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 [by the untutored native speaker]. (Rumelhart and McClelland, PDPII, p. 217)

Their own work is intended to provide "an alternative to explicit inaccessible rules ... a mechanism in which there is no explicit representation of a rule" (p. 217). The implication, or invited inference, seems to be that a formal rule is an eliminable descriptive convenience unless inscribed somewhere and examined by the neural equivalent of a read-head in the course of linguistic information processing.

In fact, there is no necessary link between realistic interpretation of rule theories and the "explicit inaccessible" view. Rules *could* be explicitly inscribed and accessed, but they *also* could be implemented in hardware in such a way that every consequence of the rule-system holds. If the latter turns out to be the case in a cognitive domain, there is a clear sense in which the rule-theory is validated—it is exactly true—rather than faced with a competing alternative or relegated to the status of an approximate convenience.[35]

Consider pattern-associators like Rumelhart and McClelland's, which gives symbolic output from symbolic input. Under a variety of conditions, it will function as a rule-implementer. To take only the simplest, suppose that all connection weights are 0 except those from the input node for feature f_i to the output node for f_i, which are set to 1. Then the network will implement the identity map. There is no read-head, write-head, or executive overseeing the operation, yet it is legitimate and even enlightening to speak of it in terms of rules manipulating symbols.

More realistically, one can abstract from the RM pattern associator an implicit theory implicating a "representation" consisting of a set of unordered Wickelfeatures and a list of "rules" replacing Wickelfeatures with other Wic-

[35] Note as well that many of the examples offered to give common-sense support to the desirability of eliminating rules are seriously misleading because they appeal to a confusion between attributing a *rule-system* to an entity and attributing the *wrong* rule-system to an entity. An example that Rumelhart and McClelland cite, in which it is noted that bees can create hexagonal cells in their hive with no knowledge of the rules of geometry, gains its intuitive force because of this confusion.

kelfeatures. Examining the properties of such rules and representations is quite revealing. We can find out what it takes to add /d/ to a stem; what it takes to reverse the order of phonemes in an input; whether simple local modifications of a string are more easily handled than complex global ones; and so on. The results we obtain carry over without modification to the actual pattern associator, where much more complex conditions prevail. The deficiencies of Wickelphone/Wickelfeature transformation are as untouched by the addition of thresholds, logistic probability functions, temperatures, and parameters of that ilk as they are by whether the program implementing the model is written in Fortran or C.

An important role of higher-level theory, as Marr for one has made clear, is to delineate the basic assumptions that lower level models must inevitably be built on. From this perspective, the high-level theory is not some approximation whose behavior offers a gross but useful guide to reality. Rather, the relation is one of embodiment: the lower-level theory embodies the higher level theory, and it does so with exactitude. The RM model has a theory of linguistic knowledge associated with it; it is just that the theory is so unorthodox that one has to look with some care to find it. But if we want to understand the model, dealing with the embodied theory is not a convenience, but a necessity, and it should be pushed as far as possible.

8.2.1. When does a network implement a rule?

Nonetheless, as we pointed out in the Introduction, it is not a logical necessity that a cognitive model implement a symbolic rule system, either a traditional or a revisionist one; the "eliminative" or rule-as-approximation connectionism that Rumelhart, McClelland, and Smolensky write about (though do not completely succeed in adhering to in the RM model) is a possible outcome of the general connectionist program. How could one tell the difference? We suggest that the crucial notion is the *motivation* for a network's structure.

In a radical or eliminative connectionist model, the overall properties of the rule-theory of a domain are not only caused by the mechanisms of the micro-theory (that is, the stipulated properties of the units and connections) but follow in a natural way from micro-assumptions that are well-motivated on grounds that have nothing to do with the structure of the domain under macro-scrutiny. The rule-theory would have second-class status because its assumptions would be epiphenomena: if you really want to understand why things take the shape they do, you must turn not to the axioms of a rule-theory but to the micro-ecology that they follow from. The intuition behind the symbolic paradigm is quite different: here rule-theory drives micro-theory; we expect to find many characteristics of the micro-level which make no micro-sense, do not derive from natural micro-assumptions or interactions, and can only be understood in terms of the higher-level system being implemented.

The RM pattern associator again provides us with some specific examples. As noted, it is surely significant that the regular past-tense morphology leaves the stem completely unaltered. Suppose we attempt to encode this in the pattern associator by pre-setting it for the identity map; then for the vast majority of items (perhaps more than 95% on the whole vocabulary), most connections will not have to be changed at all. In this way, we might be able to make the learner pay (in learning time) for divergences from identity. But such a setting has no justification from the micro-level perspective, which conduces only to some sort of uniformity (all weights 0, for example, or all random); the labels that we use from our perspective as theorists are invisible to the units themselves, and the connections implementing the identity map are indistinguishable at the micro-level from any other connections. Wiring it in is an implementational strategy driven by outside considerations, a fingerprint of the macro-theory.

An actual example in the RM model as it stands is the selective blurring of Wickelfeature representations. When the Wickelfeature ABC is part of an input stem, extra Wickelfeatures XBC and ABY are also turned on, but AXC is not: as we noted above (see also Lachter & Bever, 1988), this is motivated by the macro-principles that individual phonemes are the significant units of analysis and that phonological interactions when they occur generally involve adjacent pairs of segments. It is not motivated by any principle of micro-level connectionism.

Even the basic organization of the RM model, simple though it is, comes from motives external to the micro-level. Why should it be that the stem is mapped to the past tense, that the past tense arises from a modification of the stem? Because a sort of intuitive proto-linguistics tells us so. It is easy to set up a network in which stem and past tense are represented only in terms of their semantic features, so that generalization gradients are defined over semantic similarity (e.g. *hit* and *strike* would be subject to similar changes in the past tense), with the unwelcome consequence that no phonological relations will 'emerge'. Indeed, the telling argument against the RM pattern associator as a model of linguistic knowledge is that its very design forces it to blunder past the major generalizations of the English system. It is not unthinkable that many of the design flaws could be overcome, resulting in a connectionist network that learns more insightfully. But subsymbolism or eliminative connectionism, as a radical metatheory of cognitive science, will not be vindicated if the principal structures of such hypothetical improved models turn out to be dictated by higher-level theory rather than by micronecessities. To the extent that connectionist models are not mere isotropic node tangles, they will themselves have properties that call out for explanation. We expect that in many cases, these explanations will constitute the macro-theory of the rules that the system would be said to implement.

Here we see, too, why radical connectionism is so closely wedded to the notion of blank slates, simple learning mechanisms, and vectors of "teaching" inputs juxtaposed unit-by-unit with the networks' output vectors. If you *really* want a network not to implement *any* rules at all, the properties of the units and connections at the micro-level must suffice to organize the network into something that behaves intelligently. Since these units are too simple and too oblivious to the requirements of the computational problem that the entire network will be required to solve to do the job, the complexity of the system must derive from the complexity of the set of environmental inputs causing the units to execute their simple learning functions. One explains the organization of the system, then, only in terms of the structure of the environment, the simple activation and learning abilities of the units, and the tools and language of those aspects of statistical mechanics apropos to the aggregate behavior of the units as they respond to environmental contingencies (as in Hinton & Sejnowski, 1986; Smolensky, 1986)—the rules genuinely would have no role to play.

As it turns out, the RM model requires both kinds of explanation—implemented macrotheory and massive supervised learning—in accounting for its asymptotic organization. Rumelhart and McClelland made up for the model's lack of proper rule-motivated structure by putting it into a teaching environment that was unrealistically tailored to produce much of the behavior they wanted to see. In the absence of macro-organization the environment must bear a very heavy burden.

Rumelhart and McClelland (1986a, p. 143) recognize this implication clearly and unflinchingly in the two paragraphs they devote in their volumes to answering the question "Why are People Smarter than Rats?":

> Given all of the above [the claim that human cognition and the behavior of lower animals can be explained in terms of PDP networks], the question does seem a bit puzzling. ... People have much more cortex than rats do or even than other primates do; in particular they have very much more ... brain structure not dedicated to input/output—and presumably, this extra cortex is strategically placed in the brain to subserve just those functions that differentiate people from rats or even apes. ... But there must be another aspect to the difference between rats and people as well. This is that the human environment includes other people and the cultural devices that they have developed to organize their thinking processes.

We agree completely with one part: that the plausibility of radical connectionism is tied to the plausibility of this explanation.

8.3. On the properties of parallel distributed processing models

In our view the more interesting points raised by an examination of the RM model concern the general adequacy of the PDP mechanisms it uses, for it is this issue, rather than the metatheoretical ones, that will ultimately have the most impact on the future of cognitive science. The RM model is just one early example of a PDP model of language, and Rumelhart and McClelland make it clear that it has been simplified in many ways and that there are many paths for improvement and continued development within the PDP framework. Thus it would be especially revealing to try to generalize the results of our analysis to the prospects for PDP models of language in general. Although the past tense rule is a tiny fragment of knowledge of language, many of its properties that pose problems for the RM model are found in spades elsewhere. Here we point out some of the properties of the PDP architecture used in the RM model that seem to contribute to its difficulties and hence which will pose the most challenging problems to PDP models of language.

8.3.1. Distributed representations

PDP models such as RM's rely on 'distributed' representations: a large-scale entity is represented by a pattern of activation over a set of units rather than by turning on a single unit dedicated to it. This would be a strictly implementational claim, orthogonal to the differences between connectionist and symbol-processing theories, were it not for an additional aspect: the units have semantic content; they stand for (that is, they are turned on in response to) specific properties of the entity, and the entity is thus represented solely in terms which of those properties it has. The links in a network describe strengths of association between properties, not between individuals. The relation between features and individuals is one-to-many in both directions: Each individual is described as a collection of many features, and each feature plays a role in the description of many individuals.

Hinton et al. (1986) point to a number of useful characteristics of distributed representations. They provide a kind of content-addressable memory, from which individual entities may be called up through their properties. They provide for automatic generalization: things true of individual X can be inherited by individual Y inasmuch as the representation of Y overlaps that of X (i.e. inasmuch as Y shares properties with X) and activation of the overlapping portion during learning has been correlated with generalizable properties. And they allow for the formation of new concepts in a system via new combinations of properties that the system already represents.

It is often asserted that distributed representation using features is uniquely available to PDP models, and stands as the hallmark of a new paradigm of cognitive science, one that calculates not with symbols but with what Smolensky (in press) has dubbed 'subsymbols' (basically, what Rumelhart, McClelland, and Hinton call 'microfeatures'). Smolensky puts it this way:

> (18) Symbols and Context Dependence.
> In the symbolic paradigm, the context of a symbol is manifest around it, and consists of other symbols; in the subsymbolic paradigm, the context of a symbol is manifest inside it, and consists of subsymbols.

It is striking, then, that one aspect of distributed representation—featural decomposition—is a well-established tool in every area of linguistic theory, a branch of inquiry securely located in (perhaps indeed paradigmatic of) the 'symbolic paradigm'. Even more striking, linguistic theory calls on a version of distributed representation to accomplish the very goals that Hinton et al. (1986) advert to. Syntactic, morphological, semantic, and phonological entities are analyzed as feature complexes so that they can be efficiently content-addressed in linguistic rules; so that generalization can be achieved across individuals; so that 'new' categories can appear in a system from fresh combinations of features. Linguistic theory also seeks to make the correct generalizations inevitable given the representation. One influential attempt, the 'evaluation metric' hypothesis, proposed to measure the optimality of linguistic rules (specifically phonological rules) in terms of the number of features they refer to; choosing the most compact grammar would guarantee maximal generality. Compare in this regard Hinton et al.'s (1986, p. 84) remark about types and instances:

> ... the relation between a type and an instance can be implemented by the relationship between a set of units [features] and a larger set [of features] that includes it. Notice that the more general the type the smaller the set of units [features] used to encode it. As the number of terms in an intensional [featural] description gets smaller, the corresponding extensional set [of individuals] gets larger.

This echoes exactly Halle's (1957, 1962) observation that the important general classes of phonemes were among those that could be specified by small sets of features. In subsequent linguistic work we find thorough and continuing exploration of a symbol-processing content-addressing automatically-generalizing rule-theory built, in part, on featural analysis. No distinction-in-principle between PDP and all that has gone before can be linked to the presence or absence of featural decomposition (one central aspect of distributed representation) as the key desideratum. Features analyze the structure of paradigms—the way individuals contrast with comparable individuals—and any theory, macro, micro, or mini, that deals with complex entities can use them.

Of course, distributed representation in PDP models implies more than just featural decomposition: an entity is represented as *nothing but* the features it is composed of. Concatenative structure, constituency, variables, and their binding—in short, syntagmatic organization—are virtually abandoned. This is where the RM model and similar PDP efforts really depart from previous work, and also where they fail most dramatically.

A crucial problem is the difficulty PDP models have in representing *individuals* and *variables* (this criticism is also made by Norman, 1986, in his generally favorable appraisal of PDP models). The models represent individual objects as sets of their features. Nothing, however, represents the fact that a collection of features corresponds to an existing individual: that it is distinct from a twin that might share all its features, or that an object similar to a previously viewed one is a single individual that has undergone a change as opposed to two individual objects that happen to resemble one another, or that a situation has undergone a change if two identical objects have switched positions.[36] In the RM model, for example, this problem manifests itself in the inability to supply different past tenses for homophonous verbs such as *wring* and *ring*, or to enforce a categorical distinction between morphologically disparate verbs that are given similar featural representations such as *become* and *succumb*, to mention just two of the examples discussed in Section 4.3.

As we have mentioned, a seemingly obvious way to handle this problem—just increase the size of the feature set so that more distinctions can be encoded—will not do. For one thing, the obvious kinds of features to add, such as semantic features to distinguish homophones, gives the model too much power, as we have mentioned: it could use any semantic property or combination of semantic and phonological properties to distinguish inflectional rules, whereas in fact only a relatively small set of features are ever encoded inflectionally in the world's languages (Bybee, 1985; Talmy, 1985). Furthermore, the crucial properties governing choice of inflection are not semantic at all but refer to abstract morphological entities such as basic lexical itemhood or roothood. Finally, this move would commit one to the prediction that semantically-related words are likely to have similar past tenses, which is just not true (compare, e.g. *hit/hit* versus *strike/struck* versus *slap/slapped*

[36] We thank David Kirsh for these examples.

(similar meanings, different kinds of past tenses) or *stand/stood* versus *understand/understood* versus *stand out/stood out* (different meanings, same kind of past tense). Basically, increasing the feature set is only an approximate way to handle the problem of representing individuals; by making finer distinctions it makes it less likely that individuals will be confused but it still does not encode individuals as individuals. The relevant difference between *wring* and *ring* as far as the past tense is concerned is that they are different words, pure and simple.[37]

A second way of handling the problem is to add arbitrary features that simply distinguish words. In the extreme case, there could be a set of *n* features over which *n* orthogonal patterns of activation stand in one-to-one correspondence with *n* lexical items. This won't work, either. The basic problem is that distributed representations, when they are the *only* representations of objects, face the conflicting demands of keeping individuals distinct and providing the basis for generalization. As it stands, Rumelhart and McClelland must walk a fine line between keeping similar words distinct and getting the model to generalize to new inputs—witness their use of Wickelfeatures over Wickelphones, their decision to encode a certain proportion of incorrect Wickelfeatures, their use of a noisy output function for the past tense units, all designed to blur distinctions and foster generalization (as mentioned, the effort was only partially successful, as the model failed to generalize properly to many unfamiliar stems). Dedicating some units to representing wordhood would be a big leap in the direction of nongeneralizability. With orthogonal patterns representing words, in the extreme case, word-specific output features could be activated accurately in every case and the discrepancy between computed-output and teacher-supplied-input needed to strengthen connections from the relevant stem *features* would never occur. Intermediate solutions, such as having a relatively small set of word-distinguishing features available to distinguish homophones with distinct endings, might help. But given the extremely delicate balance between discriminability and generalizability, one won't know until it is tried, and in any case, it would at best be a hack that did not tackle the basic problem at hand: individuating individuals, and associating them with the abstract predicates that govern the permissible generalizations in the system.

The lack of a mechanism to bind sets of features together as individuals causes problems at the output end, too. A general problem for coarse-coded

[37] Of course, another problem with merely increasing the feature set, especially if the features are conjunctive, is that the network can easily grow too large very quickly. Recall that Wickelphones, which in principle can make finer distinctions than Wickelfeatures, would have required a network with more than two billion connections.

distributed representations is that when two individuals are simultaneously represented, the system can lose track of which feature goes with which individual—leading to "illusory conjunctions" where, say, an observer may be unable to say whether he or she is seeing a blue circle and a red triangle or a red triangle and a blue circle (see Hinton et al., 1986; Treisman & Schmidt, 1982). The RM model simultaneously computes past tense output features corresponding to independent subregularities which it is then unable to keep separate, resulting in incorrect blends such as *slept* as the past tense of *slip*—a kind of self-generated phonological illusory conjunction. The current substitute for a realistic binding mechanism, namely the "whole-string binding network", does not do the job, and we are given no reason to believe that a more realistic and successful model is around the corner. The basic point is that the binding problem is a core deficiency of this kind of distributed representation, not a minor detail whose solution can be postponed to some later date.

The other main problem with features-only distributed representations is that they do not easily provide *variables* that stand for sets of individuals regardless of their featural decomposition, and over which quantified generalizations can be made. This dogs the RM model in many places. For example, there is the inability to represent certain reduplicative words, in which the distinction between a feature occurring once versus occurring twice is crucial, or in learning the general nature of the rule of reduplication, where a morpheme must be simply copied: one needs a variable standing for an *occurrence* of a morpheme independent of the particular features it is composed of. In fact, even the English regular rule of adding /d/ is never properly learned (that is, the model does not generalize it properly to many words), because in essence the real rule causes an affix to be added to a "word", which is a variable standing for *any* admissible phone sequence, whereas the model associates the family of /d/ features with a list of *particular* phone sequences it has encountered instead. Many of the other problems we have pointed out can also be traced to the lack of variables.

We predict that the kind of distributed representation used in the two layer pattern-associators like the one in the RM model will cause similar problems anywhere they are used in modeling middle- to high-level cognitive processes.[38] Hinton, McClelland, and Rumelhart (p. 82) themselves provide an example that (perhaps inadvertently) illustrates the general problem:

[38] Within linguistic semantics, for example, a well-known problem is that if semantic representation is a set of features, how are propositional connectives defined over such feature sets? If P is a set of features, what function of connectionist representation will give the set for ~P?

> People are good at generalizing newly acquired knowledge. ... If, for example, you learn that chimpanzees like onions you will probably raise your estimate of the probability that gorillas like onions. In a network that uses distributed representations, this kind of generalization is automatic. The new knowledge about chimpanzees is incorporated by modifying some of the connection strengths so as to alter the causal effects of the distributed pattern of activity that represents chimpanzees. The modification automatically changes the causal effects of all similar activity patterns. So if the representation of gorillas is a similar activity pattern over the same set of units, its causal effects will be changed in a similar way.

This venerable associationist hypothesis about inductive reasoning has been convincingly discredited by contemporary research in cognitive psychology. People's inductive generalizations are not automatic responses to similarity (in any non-question-begging sense of similarity); they depend on the reasoner's unconscious "theory" of the domain, and on any theory-relevant fact about the domain acquired through any route whatsoever (communicated verbally, acquired in a single exposure, inferred through circuitous means, etc.), in a way that can completely override similarity relations (Carey, 1985; de Jong & Mooney, 1986; Gelman & Markman, 1986; Keil, 1986; Osherson, Smith, & Shafir, 1986; Pazzani, 1987). To take one example, knowledge of how a set of perceptual features was caused, or knowledge of the "kind" that an individual is an example of, can override any generalizations inspired by the object's features themselves: for example, an animal that looks exactly like a skunk will nonetheless be treated as a raccoon if one is told that the stripe was painted onto an animal that had raccoon parents and raccoon babies (see Keil, 1986; who demonstrates that this phenomenon occurs in children and is not the result of formal schooling). Similarly, even a basketball ignoramus will not be seduced by the similarity relations holding among the typical starting players of the Boston Celtics and those holding among the starting players of the Los Angeles Lakers, and thus will not be tempted to predict that a yellow-shirted blond player entering the game will run to the Celtics' basket when he gets the ball just because all previous blond players did so. (Hair color, nonetheless, might be used in qualitatively different generalizations, such as which players will be selected to endorse hair care products.) The example, from Pazzani and Dyer (1987), is one of many that have led to artificial intelligence systems based on "explanation-based learning" which has greater usefulness and greater fidelity to people's commonsense reasoning than the "similarity-based learning" that Hinton et al.'s example system performs automatically (see, e.g., de Jong & Mooney, 1986). Osherson et al. (1987), also analyse the use of similarity as a basis for generalization and show its inherent problems; Gelman and Markman (1986) show how preschool children shelve similarity relations when making inductive generalizations about natural kinds.

The point is that people's inductive inferences depend on variables assigned to sets of individuals that pick out some properties and completely ignore others, differently on different occasions, depending in knowledge-specific ways on the nature of the inductive inference to be made on that occasion. Furthermore the knowledge that can totally alter or reverse an inductive inference is not just another pattern of trained feature correlations, but depends crucially on the structured propositional content of the knowledge: learning that all gorillas are exclusively carnivorous will lead to a different generalization about their taste for onions than learning that some or many are exclusively carnivorous or that it is not the case that all gorillas are exclusively carnivorous, and learning that a *particular* gorilla who happens to have a broken leg does not like onions will not necessarily lead to *any* tendency to project that distaste onto other injured gorillas and chimpanzees. Though similarity surely plays a role in domains of which people are entirely unfamiliar, or perhaps in initial gut reactions, full-scale intuitive inference is not a mere reflection of patterns of featural similarity that have been intercorrelated in the past. Therefore one would not want to use the automatic-generalization properties of distributed representations to provide an account of human inductive inference in general. This is analogous to the fact we have been stressing throughout, namely that the past tense inflectional system is not a slave to similarity but it is driven in precise ways by speakers' implicit "theories" of linguistic organization.

In sum, featural decomposition is an essential feature of standard symbolic models of language and cognition, and many of the successes of PDP models simply inherit these advantages. However what is unique about the RM model and other two-layer pattern associators is the claim that individuals and types are represented as nothing but activated subsets of features. This impoverished mechanism is viable neither in language nor in cognition in general. The featural decomposition of an object must be available to certain processes, but can only be one of the records associated with the object and need not enter into all the processes referring to the object. Some symbol referring to the object qua object, and some variable types referring to task-relevant classes of objects that cut across featural similarity, are required.

8.3.2. Distinctions among subcomponents and abstract internal representations

The RM model collapses into a single input–output module a mapping that in rule-based accounts is a composition of several distinct subcomponents feeding information into one another, such as derivational morphology and inflectional morphology, or inflectional morphology and phonology. This, of course, is what gives it its radical look. If the subcomponents of a traditional account were kept distinct in a PDP model, mapping onto distinct subnetworks or pools of units with their own inputs and outputs, or onto distinct layers of a multilayer network, one would naturally say that the network simply implemented the traditional account. But it is just the factors that differentiate Rumelhart and McClelland's collapsed one-box model from the traditional accounts that causes it to fail so noticeably.

Why do Rumelhart and McClelland have to obliterate the traditional decomposition to begin with? The principal reason is that when one breaks a system down into components, the components must communicate by passing information—internal representations—among themselves. But because these are *internal* representations the environment cannot "see" them and so cannot adjust them during learning via the perceptron convergence procedure used in the RM model. Furthermore, the internal representations do not correspond directly to environmental inputs and outputs and so the criteria for matches and mismatches necessary to drive the convergence procedure are not defined. In other words the representations used in decomposed, modular systems are *abstract*, and many aspects of their organization cannot be learned in any obvious way. (Chomsky, 1981, calls this the argument from "poverty of the stimulus".) Sequences of morphemes resulting from factoring out phonological changes are one kind of abstract representation used in rule systems; lexical entries distinct from phonetic representations are another; morphological roots are a third. The RM model thus is composed of a single module mapping from input directly to output in part because there is no realistic way for their convergence procedure to learn the internal representations of a modular account properly.

A very general point we hope to have made in this paper is that symbolic models of language were not designed for arbitrary reasons and preserved as quaint traditions; the distinctions they make are substantive claims motivated by empirical facts and cannot be obliterated unless a new model provides equally compelling accounts of those facts. Designing a model that can record hundreds of thousands of first-order correlations can simulate some but not all of this structure and is unable to explain it or to account for the structures that do not occur across languages. Similar conclusions, we predict, will emerge from other cognitive domains that are rich in data and theory. It is unlikely that any model will be able to obliterate distinctions among subcomponents and their corresponding forms of abstract internal representations that have been independently motivated by detailed study of a domain of cognition. This alone will sharply brake any headlong movement away from the kinds of theories that have been constructed within the symbolic framework.

8.3.3. Discrete, categorical rules

Despite the graded and frequency-sensitive responses made by children and by adults in their speech errors and analogical extensions in parts of the strong verb system, many aspects of knowledge of language result in categorical judgments of ungrammaticality. This fact is difficult to reconcile with any mechanism that at asymptote leaves a number of candidates at suprathreshold strength and allows them to compete probabilistically for expression (Bowerman, 1987, also makes this point). In the present case, adult speakers assign a single past tense form to words they represent as being "regular" even if subregularities bring several candidates to mind (*e.g. brought/*brang/* bringed*); and subregularities that may have been partially productive in childhood are barred from generating past tense forms when verbs are derived from other syntactic categories (e.g. **pang; *high-stuck*) or are registered as being distinct lexical items from those exemplifying subregularities (e.g. **I broke the car*). Categorical judgments of ungrammaticality is a common (though not all-pervasive) property of linguistic judgments of novel words and strings, and cannot be predicted by semantic interpretability or any prior measure or "similarity" to known words or strings (e.g. **I put; *The child seems sleeping; *What did you see something?*). Obviously PDP models can display categorical judgments by various kinds of sharpening and threshold circuits; the question is whether models can be built—other than by implementing standard symbolic theories—in which the quantitatively strongest output prior to the sharpening circuit invariably corresponds to the unique qualitatively appropriate response.

8.3.4. Unconstrained correlation extraction

It is often considered a virtue of PDP models that they are powerful learners; virtually any amount of statistical correlation among features in a set of inputs can be soaked up by the weights on the dense set of interconnections among units. But this property is a liability if human learners are more constrained. In the case of the RM model, we saw how it can acquire rules that are not found in any language such as nonlocal conditioning of phonological changes or mirror-reversal of phonetic strings. This problem would get even worse if the set of feature units was expanded to represent other kinds of information in an attempt to distinguish homophonous or phonologically similar forms. The model also exploits subregularities (such as those of the irregular classes) that adults at best do not exploit productively (*slip/*slept* and *peep/*pept*) and at worst are completely oblivious to (e.g. lexical causatives like *sit/set—lie/lay—fall/fell—rise/raise*, which are never generalized to *cry/*cray*). The types of inflection found across human languages involves a highly constrained subset of the logically possible semantic features, feature combinations, phonological alterations, items admitting of inflection, and agreement relations (Bybee, 1985; Talmy, 1985). For example, to represent the literal meanings of the verbs *brake* and *break* the notion of a man-made mechanical device is relevant, but no language has different past tenses or plurals for a distinction between man-made versus natural objects, despite the cognitive salience of that notion. And the constrained nature of the variation in other components of language such as syntax has been the dominant theme of linguistic investigations for a quarter of a century (e.g. Chomsky, 1981). These constraints are facts that any theory of language acquisition must be able to account for; a model that can learn all possible degrees of correlation among a set of features is not a model of the human being.

8.4. Can the model be recast using more powerful PDP mechanisms?

The most natural response of a PDP theorist to our criticisms would be to retreat from the claim that the RM model in its current form is to be taken as a literal model of inflection acquisition. The RM model uses some of the simplest of the devices in the PDP armamentarium, devices that PDP theorists in general have been moving away from. Perhaps it is the limitations of these simplest PDP devices—two-layer pattern association networks—that cause problems for the RM model, and these problems would all diminish if more sophisticated kinds of PDP networks were used. Thus the claim that PDP networks rather than rules provide an exact and detailed account of language would survive.

In particular, two interesting kinds of networks, the Boltzmann Machine (Hinton & Sejnowski, 1986) and the Back-Propagation scheme (Rumelhart et al., 1986) have been developed recently that have "hidden units" or intermediate layers between input and output. These hidden units function as internal representations and as a result such networks are capable of computing functions that are uncomputable in two-layer pattern associators of the RM variety. Furthermore, in many interesting cases the models have been able to "learn internal representations". For example the Rumelhart et al. model changes not only the weights of the connections to its output units in response to an error with respect to the teaching input, but it propagates the error signal backwards to the intermediate units and changes their weights in the direction that alters their aggregate effect on the output in the right direction. Perhaps, then, a multilayered PDP network with back-propagation learning could avoid the problems of the RM model.

There are three reasons why such speculations are basically irrelevant to the points we have been making.

First, there is the gap between revolutionary manifestos and actual accomplishments. Rumelhart and McClelland's surprising claims—that language can be described only approximately by rules, that there is no induction problem in their account, and that we must revise our understanding of linguistic information processing—are based on the putative success of *their existing* model. Given that their existing model does not do the job it is said to do, the claims must be rejected. If a PDP advocate were to eschew the existing RM model and appeal to more powerful mechanisms, the only claim that could be made is that there may exist a model of unspecified design that may or may not account for past tense acquisition without the use of rules and that if it did, we should revise our understanding of language, treat rules as mere approximations, and so on. Such an assertion, of course, would have as little claim to our attention as any other claim about the hypothetical consequences of a nonexistent model.

Second, a successful PDP model of more complex design may be nothing more than an implementation of a symbolic rule-based account. The advantage of a multilayered model is precisely that it is free from the constraints that so sharply differentiate the RM model from standard ones, namely, the lack of internal representations and subcomponents. Multilayered networks, and other sophisticated models such as those that have one network that can gate the connections between two others or networks that can simulate semantic networks, production systems, or LISP primitive operations (Hinton, 1981; Touretzky, 1986; Touretzky & Hinton, 1985) are appealing because they have the ability to mimic or implement the standard operations and representations needed in traditional symbolic accounts (though perhaps with some twists). We do not doubt that it would be possible to *implement* a rule system in networks with multiple layers: after all, it has been known for over 45 years that nonlinear neuron-like elements can function as logic gates and that hence that networks consisting of interconnected layers of such elements can compute propositions (McCulloch & Pitts, 1943). Furthermore, given what we know about neural information processing and plasticity it seems likely that the elementary operations of symbolic processing will *have* to be implemented in a system consisting of massively interconnected parallel stochastic units in which the effects of learning are manifest in changes in connections. These uncontroversial facts have always been at the very foundations of the realist interpretation of symbolic models of cognition; they do not signal a departure of any sort from standard symbolic accounts. Perhaps a multilayered or gated multinetwork system could solve the tasks of inflection acquisition without simply implementing standard grammars intact (for example, they might behave discrepantly from a set of rules in a way that mimicked people's systematic divergence from that set of rules, or their intermediate layers might be totally opaque in terms of what they represent), and thus would call for a revised understanding of language, but there is no reason to believe that this will be true.

As we mentioned in a previous section, the really radical claim is that there are models that can *learn* their internal organization through a process that can be exhaustively described as an interaction between the correlational structure of environmental inputs and the aggregate behavior of the units as they execute their simple learning and activation functions in response to those inputs. Again, this is no more than a vague hope. An important technical problem is that when intermediate layers of complex networks have to learn anything in the local unconstrained manner characteristic of PDP models, they are one or more layers removed from the output layer at which discrepancies between actual and desired outputs are recorded. Their inputs and outputs no longer correspond in any direct way to overt stimuli and responses, and the steps needed to modify their weights are no longer transparent. Since differences in the setting of each tunable component of the intermediate layers have consequences that are less dramatic at the comparison stage (their effects combine in complex ways with the effects of weight changes of other units before affecting the output layer), it is harder to ensure that the intermediate layers will be properly tuned by local adjustments propagating backwards. Rumelhart et al. (1986) have dealt with this problem in clever ways with some interesting successes in simple domains such as learning to add two-digit numbers, detecting symmetry, or learning the exclusive-'or' operator. But there is always the danger in such systems of converging on incorrect solutions defined by local minima of the "energy landscape" defined over the space of possible weights, and such factors as the starting configuration, the order of inputs, several parameters of the learning function, the number of hidden units, and the innate topology of the network (such as whether all input units are connected to all intermediate units, and whether they are connected to all output units via direct paths or only through intervening links) can all influence whether the models will properly converge even in some of the simple cases. There is no reason to predict with certainty that these models will fail to acquire complex abilities such as mastery of the past tense system without wiring in traditional theories by hand—but there is even less reason to predict that they will.

These problems are exactly that, problems. They do not demonstrate that interesting PDP models of language are impossible in principle. At the same time, they show that there is no basis for the belief that connectionism will dissolve the difficult puzzles of language, or even provide radically new solutions to them. As for the present, we have shown that the paradigm example of a PDP model of language can claim nothing more than a superficial fidelity to some first-order regularities of language. More is known than just the first-order regularities, and when the deeper and more diagnostic patterns are examined with care, one sees not only that the PDP model is not a viable alternative to symbolic theories, but that the symbolic account is supported in virtually every aspect. Principled symbolic theories of language have achieved success with a broad spectrum of empirical generalizations, some of considerable depth, ranging from properties of linguistic structure to patterns of development in children. It is only such success that can warrant confidence in the reality and exactitude of our claims to understanding.

Appendix: English strong verbs

Here we provide, for the reader's convenience, an informally classified listing of all the strong verbs that we recognize in our own vocabulary (thus we omit, for example, Rumelhart & McClelland's *drag-drug*). The notation *?Verb* means that we regard *Verb* as somewhat less than usual, particularly as a strong form in the class where it's listed. The notation *??Verb* means that we regard *Verb* as obsolete (particularly in the past) but recognizable, the kind of thing one picks up from reading. The notation (+) means that the verb, in our judgment, admits a regular form. Notice that obsolescence does not imply regularizability: a few verbs simply seem to lack a usable past tense or past participle. We have found that judgments differ from dialect to dialect, with a cline of willingness-to-regularize running up from British English (south-of-London) to Canadian (Montreal) to American (general). When in doubt, we've taken the American way.

Prefixed forms are listed when the prefix-root combination is not semantically transparent.

The term 'laxing' refers to the replacement of a tense vowel or diphthong by its lax counterpart. In English, due to the Great Vowel Shift, the notion 'lax counterpart' is slightly odd: the tense-lax alternations are not *i-I, e-*ɛ, *u-U*, and so on, but rather *ay-I, i-*ɛ, *e-æ, o-*ɔ*/a, u-*ɔ*/a*. The term 'ablaut' refers to all other vowel changes.

I. T/D Superclass

1. T/D + ∅

hit, slit, split, quit, ?knit(+),[39] ?spit, ??shit, ??beshit
bid[40], rid, ?forbid
shed, spread, wed(+)[41]
let, set, beset[42], upset, wet(+)
cut, shut
put
burst, cast, cost, thrust(+)
hurt

2. T/D with laxing class

bleed, breed, feed, lead, mislead, read, speed(+), ?plead(+)
meet
hide(en), slide
bite(en), light(+), alight(+!)
shoot

3. Overt-T ending

3a. Suffix-t

burn, ??learn, ?dwell, ??spell, ???smell
?spill, ??spoil

3b. Devoicing

bend, send, spend, ?lend, ?rend
build

3c. -t with laxing

lose
deal, feel, ?kneel(+)
mean
?dream
creep, keep, leap(+), sleep, sweep(+), weep
leave

[39] *He knit a sweater* is possible, not?? *He knit his brows.*

[40] As in poker, bridge, or defense contracts.

[41] The adjective is only *wedded.*

[42] Mainly an adjective.

3d. x-ought - ought

buy, bring, catch, fight, seek, teach, think

4. Overt -D ending

4a. Satellitic laxing (cf. bleed *group)*
flee
say
hear
4b. Drop stem consonant
have
make
4c. With ablaut [ɛ - o - o]
sell, tell, foretell
4d. With unique vowel change and +n participle:
do

II. E-ɔ ablaut class

1. i/ɛ - o/ɔ - o/ɔ+n

freeze, speak, ??bespeak, steal, weave(+)[43], ?heave(+)[44]
get, forget, ??beget
??tread[45]
swear, tear, wear, ?bear, ??forbear, ??forswear

2. Satellitic x - o - o+n

awake, wake, break
choose

[43] Only in reference to carpets, etc. is the strong form possible. **The drunk wove down the road*. The adjective is *woven*.

[44] Only nautical *heave to/hove to*. **He hove his lunch*. Past participle not **hoven*.

[45] Though *trod* is common in British English, it is at best quaint in American English.

III. I - æ/ʌ - ʌ group

1. I - æ - ʌ

ring, sing, spring
drink, shrink, sink, stink
swim
begin

2. I - ʌ - ʌ

cling, ?fling, sling, sting, string, swing, wring
stick
dig
win, spin
?stink, ?slink

3. Satellites x - æ/ʌ - ʌ

run (cf. I - æ - ʌ)
hang, strike[46]
?sneak (cf. I - ʌ - ʌ)

IV. Residual clusters

1. x- u - x/o+n

blow, grow, know, throw
draw, withdraw
fly
?slay

2. e - U -e+n

take, mistake, forsake, shake, partake

[46] *Stricken* as participle as in 'from the record', otherwise as an adjective.

3. ay - aw - aw

bind, find, grind, wind

4. ay - o - X

4a. ay - o -I+n
rise, arise
write, ??smite
ride
drive, ?strive
4b. ay - o - ?
dive, shine[47]
?stride
??thrive

V. Miscellaneous

1. Pure suppletion

be
go, forgo, undergo

2. Backwards ablaut

fall, befall (cf. get–got)
hold, behold (cf. tell–told)
come, become

3. x - Y - x+n

eat
beat
see (possibly satellite of *blow*-class)
give, forgive
forbid, ??bid[48]

[47] Typically intransitive: **He shone his shoes*.

[48] As in 'ask or command to'. The past *bade* is very peculiar, *bidded* is impossible, and the past participle is obscure, though certainly not *bidden*.

4. Miscellaneous

sit, spit
stand, understand, withstand (possibly satellite of *I* - ʌ - ʌ class)
lie

5. Regular but for past participle

a. Add -n to stem (all allow -ed in participle)
sow, show, sew, prove, shear, strew
b. Add -n to ablauted stem
swell

A remark. A number of strong participial forms survive only as adjectives (most, indeed, somewhat unusual): *cleft, cloven, girt, gilt, hewn, pent, bereft, shod, wrought, laden, mown, sodden, clad, shaven, drunken, (mis)shapen.* The verb *crow* admits a strong form only in the phrase *the cock crew*; notice that *the rooster crew* is distinctly peculiar and *Melvin crew over his victory* is unintelligible. Other putative strong forms like *leant, clove, abode, durst, chid,* and *sawn* seem to us to belong to another language.

References

Anderson, J.A., & Hinton, G.E. (1981). Models of information processing in the brain. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1976). *Language, memory and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Armstrong, S.L., Gleitman, L.R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13,* 263–308.

Aronoff, M. (1976). *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Berko, J. (1958). The child's learning of English morphology. *Word, 14,* 150–177.

Bloch, B. (1947). English verb inflection. *Language 23,* 399–418.

Bowerman, M. (1987). Discussion: Mechanisms of language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Erlbaum.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Bybee, J.L. (1985). *Morphology: a study of the relation between meaning and form.* Philadelphia: Benjamins.

Bybee, J.L., & Slobin, D.I. (1982). Rules and schemes in the development and use of the English past tense. *Language, 58,* 265–289.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: Bradford Books/MIT Press.

Cazden, C.B. (1968). The acquisition of noun and verb inflections. *Child Development, 39,* 433–448.

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

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

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht, Netherlands: Foris.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper and Row.

Curme, G. (1935). *A grammar of the English language II.* Boston: Barnes & Noble.

de Jong, G.F., & Mooney, R.J. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

Ervin, S. (1964). Imitation and structural change in children's language. In E. Lenneberg (Ed.), *New directions in the study of language.* Cambridge, MA: MIT Press.

Feldman, J.A., & Ballard, D.H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205–254.

Fodor, J.A. (1968). *Psychological explanation.* New York: Random House.

Fodor, J.A. (1975). *The language of thought.* New York: T.Y. Crowell.

Francis, N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Fries, C. (1940). *American English grammar.* New York: Appleton-Century.

Gelman, S.A., & Markman, E.M. (1986). Categories and induction in young children. *Cognition, 23*, 183–209.

Gleitman, L.R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner and L.R. Gleitman (Eds.), *Language acquisition: The state of the art.* New York: Cambridge University Press.

Gordon, P. (1986). Level-ordering in lexical development. *Cognition, 21*, 73–93.

Gropen, J., & Pinker, S. (1986). Constrained productivity in the acquisition of the dative alternation. Paper presented at the 11th Annual Boston University Conference on Language Development, October.

Halle, M. (1957). In defense of the Number Two. In E. Pulgram (Ed.), *Studies presented to J. Whatmough.* Mouton: The Hague.

Halle, M. (1962). Phonology in generative grammar. *Word, 18*, 54–72.

Halwes, T., & Jenkins, J.J. (1971). Problem of serial behavior is not resolved by context-sensitive memory models. *Psychological Review, 78*, 122–29.

Hinton, G.E. (1981). Implementing semantic networks in parallel hardware. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hinton, G.E., McClelland, J.L., & Rumelhart, D.E. (1986). Distributed representations. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: Bradford Books/MIT Press.

Hinton, G.E., & Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: Bradford Books/MIT Press.

Hoard, J., & C. Sloat (1973). English irregular verbs. *Language, 49*, 107–20.

Hockett, C. (1942). English verb inflection. *Studies in Linguistics, 1.2.*, 1–8.

Jespersen, O. (1942). *A modern English grammar on historical principles, VI.* Reprinted 1961: London: George Allen & Unwin Ltd.

Keil, F.C. (1986). The acquisition of natural kinds and artifact terms. In W. Demopoulos & A. Marras (Ed.), *Language learning and concept acquisition: Foundational issues.* Norwood, NJ: Ablex.

Kiparsky, P. (1982a). From cyclical to lexical phonology. In H. van der Hulst, & N. Smith (Eds.). *The structure of phonological representations.* Dordrecht, Netherlands: Foris.

Kiparsky, P. (1982b). Lexical phonology and morphology. In I.S. Yang (Ed.), *Linguistics in the morning calm.* Seoul: Hansin, pp. 3–91.

Kucera, H., & N. Francis (1967). *Computational analysis of present-day American English.* Providence: Brown University Press.

Kuczaj, S.A. (1976). Arguments against Hurford's auxiliary copying rule. *Journal of Child Language, 3*, 423–427.

Kuczaj, S.A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16*, 589–600.

Kuczaj, S.A. (1978). Children's judgments of grammatical and ungrammatical irregular past tense verbs. *Child Development, 49*, 319–326.

Kuczaj, S.A. (1981). More on children's initial failure to relate specific acquisitions. *Journal of Child Language, 8*, 485–487.

Lachter, J., & Bever, T.G. (1988). The relation between linguistic structure and associative theories of language learning—A constructive critique of some connectionist learning models. *Cognition, 28*, 195–247 this issue.

Lakoff, G. (1987). Connectionist explanations in linguistics: Some thoughts on recent anti-connectionist papers. Unpublished electronic manuscript, ARPAnet.

Levy, Y. (1983). The use of nonce word tests in assessing children's verbal knowledge. Paper presented at the 8th Annual Boston University Conference on Language Development, October, 1983.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language sound structure.* Cambridge, MA: MIT Press.

MacWhinney, B., & Snow, C.E. (1985). The child language data exchange system. *Journal of Child Language, 12*, 271–296.

MacWhinney, B., & Sokolov, J.L. (1987). The competition model of the acquisition of syntax. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale. NJ: Erlbaum.

Maratsos, M., Gudeman, R., Gerard-Nogo, P., & de Hart, G. (1987). A study in novel word learning: The productivity of the causative. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Erlbaum.

Maratsos, M., & Kuczaj, S.A. (1978). Against the transformationalist account: A simpler analysis of auxiliary overmarkings. *Journal of Child Language, 5*, 337–345.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

McCarthy, J., & Prince, A. (forthcoming). *Prosodic morphology.*

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159–188.

McClelland, J.L., Rumelhart, D.E., & Hinton, G.E. (1986). The appeal of parallel distributed processing. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: Bradford Books/MIT Press.

McClelland, J.L., Rumelhart, D.E., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: Bradford Books/MIT Press.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Mencken, H. (1936). *The American language.* New York: Knopf.

Minsky, M. (1963). Steps toward artificial intelligence. In E.A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Newell, A., & Simon, H. (1961). Computer simulation of human thinking. *Science, 134*, 2011–2017.

Newell, A., & Simon, H. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D.A. (1986). Reflections on cognition and parallel distributed processing. In J.L. McClelland, D.E. Rumelhart, & The PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: Bradford Books/MIT Press.

Osherson, D.N., Smith, E.E., & Shafir, E. (1986). Some origins of belief. *Cognition, 24*, 197–224.

Palmer, H. (1930). *A grammar of spoken English on a strictly phonetic basis.* Cambridge: W. Heffer.

Pazzani, M. (1987). Explanation-based learning for knowledge-based systems. *International Journal of Man-Machine Studies, 26*, 413–433.

Pazzani, M., & Dyer, M. (1987). A comparison of concept identification in human learning and network learning with the generalized delta rule. Unpublished manuscript, UCLA.

Pierrehumbert, J., & Beckman, M. (1986). Japanese tone structure. Unpublished manuscript, AT&T Bell Laboratories, Murray Hill, NJ.

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

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S., Lebeaux, D.S., & Frost, L.A. (1987). Productivity and conservatism in the acquisition of the passive. *Cognition, 26*, 195–267.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind: A symposium*. New York: NYU Press.

Pylyshyn, Z.W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: Bradford Books/MIT Press.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal representation of categories. *Cognitive Psychology, 7*, 573–605.

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Spartan.

Ross, J.R. (1975). Wording up. Unpublished manuscript, MIT.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1986a). PDP models and general issues in cognitive science. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1986b). On learning the past tenses of English verbs. In J.L. McClelland, D.E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Rumelhart, D.E., McClelland, J.L., and the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Sampson, G. (1987). A turning point in linguistics. *Times Literary Supplement*, June 12, 1987, 643.

Savin, H., & Bever, T.G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior, 9*, 295–302.

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-ordering mechanism in sentence production. In W.E. Cooper & E.C.T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum.

Sietsema, B. (1987). Theoretical commitments underlying Wickelphonology. Unpublished manuscript. MIT.

Sloat, C., & Hoard, J. (1971). The inflectional morphology of English. *Glossa, 5*, 47–56.

Slobin, D.I. (1971). On the learning of morphological rules: A reply to Palermo and Eberhart. In D.I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium*. New York: Academic Press.

Slobin, D.I. (1985). Crosslinguistic evidence for the language-making capacity. In D.I. Slobin (Ed.), *The crosslinguistic study of language acquisition. Volume II: Theoretical issues*. Hillsdale, NJ: Erlbaum.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Smolensky, P. (in press). The proper treatment of connectionism. *Behavioral and Brain Sciences.*

Sommer, B.A. (1980). The shape of Kunjen syllables. In D.L. Goyvaerts (Ed.), *Phonology in the 80's*. Ghent: Story-Scientia.

Sweet, H. (1892). *A new English grammar, logical and historical.* Oxford: Clarendon.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and semantic description, Vol. 3: Grammatical categories and the lexicon*. New York: Cambridge University Press.

Touretzky, D. (1986). BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.*

Touretzky, D., & Hinton, G.E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.*

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*, 107–141.

van der Hulst, H., & Smith, N. (Eds.) (1982). *The structure of phonological representations*. Dordrecht, Netherlands: Foris.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition.* Cambridge, MA: MIT Press.

Wickelgren, W.A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76*, 1–15.

Williams, E. (1981). On the notions "lexically related" and "head of a word". *Linguistic Inquiry, 12*, 245–274.

Résumé

La connaissance du langage repose-t-elle sur la représentation mentale de règles? Rumelhart et McClelland ont développé un modèle connectioniste (parallel distributed processing, PDP) de l'acquisition du passé anglais qui parvient à produire la forme passé d'un certain nombre de verbes, à la fois réguliers (*walk/walked*) et irréguliers (*go/went*), à partir de leurs racines, et qui semble commettre certaines des erreurs et passer par certains des étapes de développement des enfants qui apprennent le passé anglais. Pourtant, le modèle ne contient pas de règles explicites; il est exclusivement constitué d'un ensemble d'unités qui représentent des trigrammes de traits phonétiques de la racine, d'un ensemble d'unités qui représentent des trigrammes de traits phonétiques de la forme passée de la racine, et d'un réseau de connections entre les deux ensembles d'unités, connections dont la force varie en fonction de l'apprentissage. La conclusion de Rumelhart & McClelland est que les règles linguistiques ne sont peut-être en fait que des approximations pratiques et que les processus causaux réels de l'utilisation et de l'acquisition du langage doivent être caractérisés en termes de transfert de niveaux d'activation entre unités et de modification du poids de leurs connections. Nous avons analysé en détail les hypothèses linguistiques et de développement qui sous-tendent leur modèle et avons découvert que (1) le modèle ne peut pas représenter certains mots, (2) il ne peut pas apprendre beaucoup de règles, (3) il peut apprendre des règles que l'on ne rencontre dans aucune langue humaine, (4) il ne peut pas expliquer certaines régularités morphologiques et phonologiques, (5) il ne peut pas expliquer les différences entre formes régulières et irrégulières, (6) il ne parvient pas à accomplir la tâche qui lui a été assigné, à savoir apprendre le passé anglais, (7) il explique incorrectement deux phénomènes de développement: les étapes de sur-régularisation de formes irrégulières comme *bringed*, et l'apparition de formes doublement marquées comme *ated*, enfin, (8) il donne une explication de deux autres phénomènes (la surrégularisation peu fréquente des verbes qui se terminent en *t/d*, et l'ordre d'acquisition des différentes sous-classes irrégulières) qui est indifférenciable de celle fournie par des théories utilisant des règles. En outre, nous montrons que c'est l'architecture connectioniste du modèle qui est responsable de ses nombreux défauts. Notre conclusion est que les affirmations des connectionistes quant à l'inutilité des règles dans les explications doivent être rejetées et que, bien au contraire, toutes les données militent en faveur de l'existence de telles règles.