Rules and connections in human language

Alan Prince and Steven Pinker

Recently 'connectionist' or 'parallel distributed processing' (PDP) approaches to brain modelling have attracted an enormous amount of attention. These models are said to be faithful to neurophysiological and to behavioral data in a way that previous approaches based on symbolic computation were not. A PDP simulation by Rumelhart and McClelland of children's acquisition of the past tense in English has been one of the most famous demonstrations of the advantages of the connectionist approach. In a recent special issue of the journal Cognition *devoted to Connectionism and Symbol Systems, Steven Pinker and Alan Prince examine this model and the relevant data in great detail, finding severe limitations in the ability of current PDP models to explain human language and cognition. The key points of their analysis are summarised in the following article.*

Alan Prince is at the Program in Linguistics and Cognitive Science, Brandeis University, Waltham, MA, USA, and Steven Pinker is at the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.

Everyone hopes that the discoveries of neuroscience will help explain human intelligence, but no one expects such an explanation to be done in a single step. Neuroscience and cognitive science, it is hoped, will converge on an intermediate level of 'cognitive architecture', which would specify the elementary information processes that arise as a consequence of the properties of neural tissue and that serve as the building blocks of the cognitive algorithms that execute intelligent behavior. This middle level has proven to be elusive. Neuroscientists study firing rates, excitation, inhibition, plasticity; cognitive scientists study rules, representations, symbol systems. Although it's relatively easy to imagine ways to run cognitive symbol systems on digital computers, how they could be implemented in neural hardware has remained obscure. Any theory of this middle level faces a formidable set of criteria: it must satisfy the constraints of neurophysiology and neuroanatomy, yet supply the right kind of computational power to serve as the basis for cognition.

Recently there has been considerable enthusiasm for a theory that claims to do just that. Connectionist or Parallel Distributed Processing (PDP) models try to model cognitive systems using networks of large numbers of densely interconnected units. The units transmit signals to one another along weighted connections; they 'compute' their output signals by weighting each of their input signals by the strength of the connection it comes in on, summing the weighted inputs, 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 and a 'desired' output provided by a set of 'teaching' inputs[1,2]. These are not meant to be genuine neural models; although some of their properties are reminiscent of the nervous system, others, such as the teaching and learning mechanisms, have no neural analogue, and much of what we know of the topology of neural connectivity plays no role[3]. However, their proponents refer to them as 'brain-style' or 'brain-metaphor' models, and they have attracted enormous interest among neuroscientists[4]. Much of this interest comes from demonstrations that show how the models can exhibit rule-like behavior without containing rules. The implication is that PDP networks eventually might be consistent with both neurophysiology and with a revised, but adequate theory of cognition, providing the long-sought bridge.

The most dramatic and frequently-cited demonstration of the rule-like behavior of PDP systems comes from a model of the acquisition of the past tense in English[5]. It addresses a phenomenon that has served as a textbook example of the role of rules in cognitive behavior[6]. Young children use regular ('walked') and irregular ('broke') verbs early on, but then begin to generalize the regular 'ed' ending, saying 'breaked' and considerably later, 'broked' as well. By kindergarten they can convert a nonsense word 'jick' provided by an experimenter into 'jicked', and easily differentiate the three different phonological variants of the regular suffix: 't' for words ending in an unvoiced consonant ('walked'), 'd' for words ending in a voiced phoneme ('jogged'), or 'ed' for words ending in a 't' or 'd' ('patted'). According to the traditional explanation of the developmental sequence, children first memorize past forms directly from their parents' speech, then coin a rule that generates them productively.

Remarkably, Rumelhart and McClelland's network model exhibits the same general type of behavior (and also several other developmental phenomena), but has no rules at all (see Box 1). It has no representations of words, regular versus irregular cases, roots, stems, or suffixes. Rather, it is a simple two-layer network, with a set of input units that are turned on in patterns that correspond to the verb stem, a set of output units that are turned on in patterns that correspond to the verb's past tense form, and connections between every input unit and every output unit. All that happens in learning is that the network compares its own version of the past tense form with the correct version provided by a 'teacher', and adjusts the strengths of the connections and the thresholds so as to reduce the difference. Rumelhart and McClelland suggest, and many are quick to agree, that this shows the viability of associationist theories of language acquisition, despite their virtual abandonment by linguists 25 years ago[7]. A system can show rule-like behavior without actually containing rules; perhaps the more sophisticated PDP version of associationism can serve as the basis of a revised theory of the psychology of language at the same time as its underlying mechanisms are tuned to be more faithful to neurophysiology.

0378 - 5912/88/$02.00

Box 1. How the Rumelhart–McClelland model works

Rumelhart and McClelland's model, in its trained state, should take any stem as input and emit the corresponding past tense form. 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. (In English, the stem is the same as the infinitive and the uninflected present tense). A graphical representation of the model is shown in this box; of its three components, the center one, the 'pattern associator', is the most important theoretically.

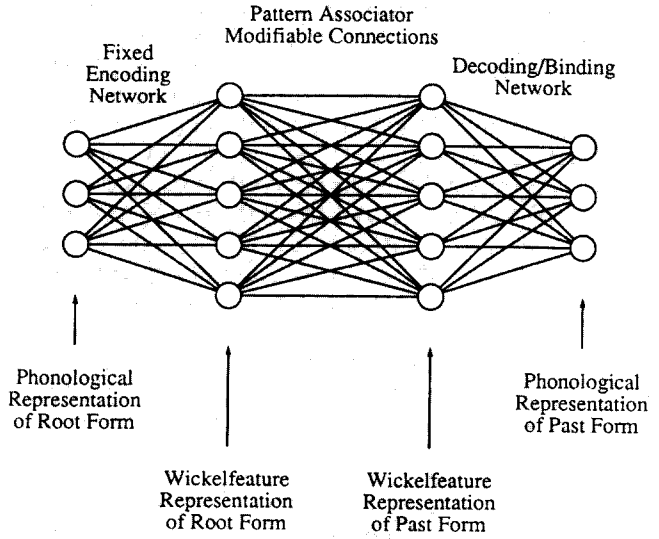
The model's pattern associator is a simple network with two layers of nodes (or 'units'), one layer for representing input, the other for output. Nodes in the R–M model may only be 'on' or 'off'; each node therefore represents a single binary feature, 'off' or 'on' marking the absence or presence of a certain property that a word may have. Every distinct stem must be encoded as a unique subset of input nodes; every distinct past tense form as a unique subset of output nodes.

Here a problem arises. The natural assumption would be that words are concatenations of phonemes, strings on an alphabet. But the pattern association network must analyse inputs as an unordered set of properties (codable as a set of turned-on units). Dedicating each unit to a phoneme would obliterate information about serial order, leading to the confusion of 'pit' and 'tip', 'cat' and 'tack', and so on. To overcome this problem, Rumelhart and McClelland turn to a scheme proposed by Wickelgren[20], according to which a string is represented as the set of the trigrams (3-character sequences) that it contains. (In order to mark word-edges, it is necessary to assume that word-boundary (#) is a character in the underlying alphabet.) Rumelhart and McClelland call such trigrams 'Wickelphones'. For example, the word 'strip' contains the following assortment of Wickelphones: {ip#, rip, str, tri, #st}; and 'strip' is uniquely reconstructible from this trigram set. Although certain trigram sets are consistent with more than one string, in particular those containing three Wickelphones ABC, BCA, and CAB, all words in their sample were uniquely encoded.

However, the Wickelphone itself is not suitable for the task at hand, for two compelling reasons. First, the number of possible Wickelphones for their representation of English would have multiplied out to over 43 000 nodes for the input stem, 43 000 nodes for the output past form, and two billion connections between them, too many to handle in present-day computers. Second, an interesting model must be able to generalize beyond the word set given to it, and provide past tense forms for new stems based on their similarities to the ones trained. Since phonological regularities (like those involved in past tense formation) do not treat phonemes as atomic, unanalysable wholes but pertain instead to their constituent phonetic properties like place and manner of articulation, voicing, the height and tenseness of vowels, stridency (noisiness), and so on, it is necessary that such fine-grained information be represented in the network. The Wickelphone is too coarse to support the necessary generalization. For example, any English speaker who had to provide a past tense form for the hypothetical verb 'to Bach' (as in the composer) would pronounce it bach-t, not bach-d or bach-ed, even though ch is not a consonant they have heard used in an English verb. This reaction is based on the similarity of ch to p, k, s, and so on, all of which share the feature 'unvoiced'; and the tacit knowledge that verbs ending in unvoiced consonants have their past tense suffix pronounced as t. Since atomic symbols for phonemes do not represent this similarity, decomposition of phonemes into features is a standard tool of phonology. Rumelhart and McClelland therefore assume a phonetic decomposition of segments into features which are in broad outline like those used by phonologists. Rather than dedicating nodes to Wickelphones, they use what they call 'Wickelfeatures', where a Wickelfeature is a trigram of features, one from each of the 3 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'. By simplifying the Wickelfeature set in a number of other ways that will not be discussed here, Rumelhart and McClelland pared down the number of Wickelfeature nodes to 460. (See Refs 5 and 8 for a complete description of the model.)

Each word is represented by a collection of turned-on nodes that correspond to its Wickelfeatures. This gives a 'distributed' representation: an individual word does not register on its own node, but is analysed as an ensemble of properties, Wickelfeatures. As the figure shows, an "encoder" of an unspecified nature is invoked to convert an ordered phonetic string into a set of activated Wickelfeature units.

In the pattern associator, every input node is connected to every output node, so that each input Wickelfeature can influence any Wickelfeature in the output set. Because of this, the



Rumelhart–McClelland model of English past tense acquisition

device can record an immense variety of correlations between patterns of activation. How is an output generated? Suppose that a set of input nodes is turned on. Any given output node receives signals from the nodes in the active input set. The strength of each such signal is determined by a weight attached to the link along which the signal travels. The output node adds up the strengths of all incoming signals and compares this sum to a threshold value; if the sum exceeds the threshold, the node can turn on. The decision is made probabilistically: the farther the weighted sum is above (below) the threshold, the more likely the node is to turn on (off).

The untrained pattern associator starts out with no preset relations between input and output nodes: the link weights are all zero. Training involves presenting the network with an input stem and comparing the output pattern actually obtained with the desired pattern, which is provided to the network by a 'teacher' as a distinct kind of 'teaching' input (not shown in the figure). 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.

The comparison between the output pattern obtained from the network's internal computation 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 theshold 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'[21] 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 R–M 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. 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.

The structure of the encoding and decoding networks are not the focus of Rumelhart and McClelland's efforts, but they must have several special properties for the model to work properly. The input encoder was 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. Because this 'blurring' is highly selective, the process that gives rise to it cannot be interpreted as random noise. Rather, the blurred representation is intended to foster a further kind of generalization of the right sort; blurring the input representations makes the connection weights in the R–M model less likely to be able to exploit the idiosyncrasies of the words in the training set and hence reduces the model's tendency towards 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 characterizing properties of the predicted past tense form. Nothing in the model ensures that the set of activated output units will fit together to describe a legitimate word or even a unique, consistent, and well-formed string of phonemes. Since the output Wickelfeatures virtually never define such a string exactly, there is no clear sense which one knows which word (if any) the output Wickelfeatures are defining. A special mechanism called the 'Whole-String Binding Network', was programmed to provide an estimate of the model's tendencies to output possible words. Basically, this network had one unit stand for every possible string of phonemes less than 20 phonemes long (obviously, the actual set had to be pruned considerably from this astronomically large number of possibilities; this was done with the help of another mechanism that will not be discussed here). Once the set of Wickelfeature units in the past tense vector is activated, the word-like nodes in the whole-string binding network 'compete' for them. Each whole-string 'word' unit has a transient strength value that increases with the number of activated Wickelfeatures that its associated 'word' uniquely contains. ('Credit' for activated Wickelfeatures contained in several words is split among the units standing for those words.) Conversely, activated Wickelfeatures that are not contained in a word cause the strength of that word's unit to diminish. Strings whose units exceed a threshold level of strength after this competition process stabilizes are interpreted as the final output of the model.

In sum, the R–M model works as follows. The phonological string is exchanged 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 an output word by a whole-string binding network.

Of course, the fact that a computer model behaves intelligently without rules does not show that humans lack rules, any more than a wind-up mouse shows that real mice lack motor programs. Recently, a set of papers has argued that the most prominent PDP models of language are incorrect on empirical grounds[8–10]. The evidence comes from a number of sources: the nature of children's language, as observed both in experiments and naturalistic studies; regularities in the kinds of words and sentences people judge to be natural-sounding or ill-formed in their colloquial speech; and the results of the simulation runs of the models themselves. If true, the implications are important, for they bear on the claims that associative networks can explain human rule-governed intelligence. We review here the most prominent evidence, which falls into three groups: the design of the model, its asymptotic performance (which ought to approximate adult's command of everyday English), and its child-like intermediate behavior.

Evidence for the linguistic constructs lacking from the Rumelhart–McClelland model

The Rumelhart-McClelland model owes its radical look to the fact that it has nothing corresponding to the formal linguistic notions 'segment', 'string', 'stem', 'affix', 'word', 'root', 'regular rule', or 'irregular exception'. However, in standard psycholinguistic theories these entities are not mere notational conveniences, but constructs designed to explain facts about the organization of language. By omitting the constructs without adequate substitutes, the R–M model is inconsistent with these facts.

Strings and segments

According to standard theories, a word's phonological representation contains a string of segments (phonemes), each segment decomposed into features that correspond to aspects of the articulation or sound of the segment (e.g. voiced/unvoiced, nasal/oral, front/back). Rumelhart and McClelland, in contrast, use a completely 'distributed' representation[2], in which a word is a (simultaneous) pattern of activation over a single vector of units. This leads to an immediate problem for them: representing linear order. If each unit simply represented a phoneme or feature, the model would not be able to distinguish words in which the same sounds appear in different orders, for example 'apt', 'pat', and 'tap'. Thus Rumelhart and McClelland are led to use context-sensitive units, each of which encodes the presence of a substring of three adjacent phonological features in a word. For example, 'unvoiced-unvoiced-voiced' and 'fricative-stop-low_vowel' are two of the context-sensitive features activated for the word 'stay'. The input and output vectors each consist of 460 of these units; by activating subsets of them, it is possible to define unique patterns for the common English verbs.

There is good evidence that people don't use context-sensitive units of this kind, however. First, trisegmental units cannot uniquely encode all linguistic strings: though such units may work for English, they won't work generally. For example, the Australian language Oykangand contains distinct words 'algal' and 'algalgal'. These decompose into the very same set of context-sensitive features, and hence the model is incapable of distinguishing them. Second, the features make the wrong predictions about psychological similarity. Pairs of strings that differ in terms of the order of two phonemes, such as 'slit' and 'silt', are judged to sound similar, and indeed confusions among them are the probable cause of certain changes in the history of English such as 'brid' to 'bird' or 'thrid' to 'third'. However, if the atomic units of description correspond to (what we usually think of as) triples, then <abc> and <acb>, as atoms, are entirely distinct (one mustn't be misled by the fact that we, the theorists, use three-letter mnemonic abbreviations for them). Without introducing arbitrary tricks into the model, it is impossible to account for perceived similarities between words defined by them[8,9]. Third, the model makes the wrong prediction about the kinds of rules that should be easy to learn, hence prevalent in languages, and those rules that should be absent from languages. It is as easy for the model to learn bizarre, cross-linguistically non-existent rules for forming the past tense (such as reversing the order of the phonemes of the stem, which involves the simple association of each input unit <abc> with the output unit <cba>; or changing every phoneme to the next one in English alphabetical order; or adding a 'g' to end of a word if it begins with 'st' but a 'p' if it begins in 'sk') as it is to learn common rules (e.g. do nothing to the stem; add a 'd' to the stem[8]).

The basic problem is that in their simple associationist architecture, the same units must represent both the decomposition of a string into phonetic components and the order in which the components are concatenated. These are conflicting demands and ultimately the units can satisy neither successfully. The Rumelhart–McClelland representational system is a case study in the difficulty of meeting the known constraints on cognitive structure. The actual units of phonological structure – from phonetic features, to segments, to syllables and stress-groups – are reasonably well-understood. Abandoning them in favor of a unit – the feature triplet – that demonstrably has no role in linguistic processes is sure to lead to major empirical problems.

Morphology and phonology

The R–M model computes a one-step mapping from the phonological features of the stem to the phonological features of the past tense form. This allows it to dispense with many of the rules and abstract representations one finds in familiar theories of language. But there is overwhelming evidence that the mapping is actually computed in several layers. Consider the pattern of differences in the suffixes in 'walked', 'jogged', and 'patted', which are contingent on the last phoneme of the stem. These differences are not unique to the past tense form: they also occur in the passive participle ('he was kicked', 'he was slugged', 'he was patted') and in adjectives ('sabre-toothed', 'long-nosed',

'one-handed'). They also occur with different suffixes altogether, such as the plural ('hawks', 'dogs', 'hoses') or the possessive ('Pat's', 'Fred's', 'George's'). They even occur in simple words lacking inflection: there are words like 'ax' and 'act', with two unvoiced consonants in a row, and words like 'adze', with two voiced consonants in a row, but no words pronounced like 'acd' or 'agt', with an unvoiced consonant followed by a voiced consonant or vice-versa. The obvious explanation is that the t-d-ed pattern has nothing to do with the past tense at all; it belongs to a different system – phonology – that adjusts words and strings so as to conform to the sound pattern of English, regardless of how the words or strings were formed. (Basically, the phonological rules here force consonant clusters at the ends of words to be either consistently voiced or consistently unvoiced, and they insert a vowel between adjacent consonants if they are too similar). The regular past tense pattern, belonging to the 'morphological' system, is simply that /d/ gets added to the end of a verb; the threefold variation is handled by a different phonological component. By collapsing the distinction into a single component, the model cannot account for the fact that the pattern of threefold variation follows from general constraints on the language as a whole.

Stem and affix

Linguistic processes tend to 'copy' stems with only minor modifications: 'walk/walked' is a pervasive pattern; 'go/went' is extremely rare. In some languages, the stem is copied twice, a phenomenon called 'reduplication': the past of 'go' would be 'gogo'. Similarly, the identity of an affix tends to be preserved across its variants: the endings for 'jog' and 'pat' are 'd' and 'ed', respectively, not 'd' and 'ob' or 'iz' and 'gu'. A subtle but important property of network models is that there is no such thing as pure copying, just modifiable connections between one set of units and another set. Only the consistency of the pairings among units can affect the operation of the model, not what the units stand for (the labels next to the units are visible to the theorist, but not to the model). Hence the prevalence of copying operations in linguistic mappings is inexplicable; the network model could just as easily learn rules that change all a's to e's, all b's to c's, and so on.

Lexical items

In standard psychological theories, a word has an 'entry' in a 'mental lexicon' that is distinct from its actual sound. This is necessary because of homophones such as 'ring' and 'wring' or 'lie' (prevaricate) and 'lie' (recline). Crucially, homophones can have different past tense forms, for example, 'rang' and 'wrung', or 'lied' and 'lay'. The R–M model, because it simply maps from phonological units to phonological units, is capable of handling such words.

A natural reaction to this phenomenon might be to suppose that the past tense form is associated with meaning as well as with sound; perhaps the different semantic feature representations of the meanings of 'ring' and 'wring' can be directly associated with their different past tense forms. Somewhat surprisingly, it turns out that meaning is almost completely irrelevant to the past tense form; such forms are sensitive to distinctions at a level of representation at which verb roots are distinct but meaningless symbols. For example, verbs like 'come, go, do, have, set, get, put, stand . . . ' each have dozens of meanings, especially in combination with 'particles' like 'in, out, up' and 'off', but they have the same irregular past tense forms in each of these semantic incarnations. This even occurs when these stems appear in combination with meaningless prefixes – 'stood/understood', 'get/forget', 'come/overcome'. (Though the prefixes are meaningless, they must be real prefixes, appearing in other words: 'overcome' and 'become', which contain intuitively recognizable prefixes, are transformed to 'overcame' and 'became', but 'succumb', which sounds similar but lacks a genuine prefix, is not transformed into 'succame'). Conversely, synonyms need not have the same kind of past tense forms: compare 'hit/hit' versus 'strike/struck' versus 'slap/slapped', which have similar meanings, but different kinds of past tenses. Thus the similarity space relevant to the irregular past tenses has no semantic dimensions in it; all that matters is gross distinctness – '*wring* is not the same word as *ring*' – not actual meaning.

Even the distinction between a 'verb' and a 'verb root' is psychologically significant. Somewhat to the puzzlement of non-scientific prescriptive grammarians, people find it natural to say 'broadcasted', not 'broadcast', 'joy-rided', not 'joy-rode', 'grandstanded', not 'grandstood', 'high-sticked' (in ice hockey) not 'high-stuck'. The reason is that 'irregularity' is a property attached to verb roots, not verbs. For each of these verbs, speakers have a sense, usually unconscious, that they were derived from nouns ('a joy-ride', 'a high-stick', etc.). Since it makes no sense for a noun to be marked in a person's mental dictionary as having an irregular 'past tense form', any verb that is felt to be derived from nouns or adjectives automatically becomes regular, hence 'joy-rided'.

What all these examples suggest is that the mental processes underlying language are sensitive to a system of representation – traditionally called 'morphology' – at which there are lawful regularities among entities that are neither sounds nor meanings, specifically, lexical items, stems, affixes, roots, and parts-of-speech.

Regular versus irregular pasts

A revolutionary aspect of the Rumelhart–McClelland model is that the regular and irregular past tense alternations are collapsed into a single network. This is an example of one of the frequently claimed advantages of connectionist systems in general: that rule-governed cases, partially rule-governed cases, and isolated exceptions are all treated uniformly[11]. But of course this is only an advantage if people show no clear-cut distinction between rule-governed and exceptional behavior.

In the case of the past tense system, however, qualitative differences can be documented. Consider these four:

(1) Irregular verbs cluster into 'family resemblance groups' that are phonologically similar: ('blow/blew' 'grow/grew', 'throw/threw') ('take/took', 'shake/shook') ('sting/stung', 'fling/flung', 'stick/stuck'). Regular verbs have nothing in common phonologically; any string can be a regular verb.

(2) Irregular pasts can be fuzzy in their naturalness or acceptability, depending on how similar they are to the central tendency of a cluster: 'wept', 'knelt', 'rent', and 'shod' sound stilted to many speakers, especially speakers of American English. In the extreme case, irregular past tense forms can sound totally bizarre: 'Last night I forwent the pleasure of grading papers' or 'I don't know how she bore it' have a very strange sound to most ears. In contrast, regular verbs, unless they are similar to an irregular cluster, have no gradient of acceptability based on their phonology: even phonologically unusual stems such as 'genuflect' yield past tense forms that sound as natural in the past tense as they are in the present tense; 'She eked out a living' is no worse-sounding than 'she ekes out a living'; 'They prescinded' no worse than 'They prescind' (even if one has no idea what 'prescind' means).

(3) There are no sufficient conditions for a verb to be in any irregular class: although 'blow' becomes 'blew' in the past, 'flow' becomes 'flowed'; although 'ring' becomes 'rang', 'string' becomes 'strung' and 'bring' becomes 'brought'. In contrast, a sufficient condition for a verb to be regular is that it not be irregular; if it is regular, its past tense form is 100% predictable.

(4) Most of the irregular alternations can only apply to verbs with a certain structure: the pattern in 'send/sent', namely to change a 'd' to a 't', requires that there be a 'd' in the stem to begin with. The regular rule, which adds a 'd' to the stem, regardless of what the stem is, can cover all possible cases by its very nature.

These differences, though subtle, all point to the same conclusion. There is a psychologically significant difference between regular and irregular verbs: the former seem to be governed by an all-or-none process – a rule – that applies across the board except where specifically pre-empted by the presence of an irregular past tense form; the latter consist of several memorized lists of similar-sounding words forming fuzzy family resemblance classes.

The model's degree of success

Despite the optimistic claims of success for the model, its actual performance is limited in significant ways. After 80000 training trials (about 200 presentations each of 420 pairs of verb stems and their correct past tense forms), the model was given 72 new verbs in a test of its ability to generalize. It made errors on 33% of these verbs. In some cases, it emitted no response at all; in others, it offered a single incorrect form; in still others, it offered both a correct and an incorrect form, unable to decide between them (these cases must count as errors: a crucial aspect of the psychology of language is that irregular forms pre-empt regular ones in people's speech – not only do people say 'went' and 'came', but they avoid saying 'goed' and 'comed').

The model's errors can be traced to several factors. First, it associates past tense features with specific stem features; it has no concept of an abstract entity 'stem' independent of the features it is composed of. Hence, if there are gaps in the phonological space defined by the stems that the model was trained on, it could fail to generalize to newly presented items occupying those gaps and emit no response at all even to common words such as 'jump' or 'warm'. Second, the model soaks up any degree of regularity in the training set, leading it to overestimate the generality of some of the vowel changes found among English irregular verbs, and resulting in spurious overregularizations such as 'shipped' as the past of 'shape' or 'brawned' as the past of 'brown'. Third, the model has no way of keeping track of separate competing responses, such as 'type' and 'typed'; hundreds of mutually incompatible features associated with a stem are all activated at once in the output feature vector, with no record of which ones cohere as target responses. Though Rumelhart and McClelland constructed a temporary, separate response-competition module to extract a cohesive response, the module could not do so effectively, producing blended hybrids such as 'typeded' for 'typed', 'membled' for 'mail', and 'squakt' for 'squat'.

Children's language

The most dramatic aspect of the past tense model is its apparent ability to duplicate the stages that children pass through: first using 'ate', then both 'ate' and 'eated' (and occasionally 'ated'), finally 'ate' exclusively. This is especially surprising because nothing changes in the model itself; it just responds passively to the teacher's input.

For this reason, it turns out that the model's changes are caused by changes in its input. Rumelhart and McClelland note that high-frequency verbs tend to be irregular and vice versa. Children, they reasoned, are likely to learn a few high-frequency verbs first, then a large number of verbs, of which an increasing proportion would probably be regular. Hence in simulating children with their model, they defined two stages. In the first stage they fed in ten high-frequency verbs (two regular and eight irregular), paired with their correct past tense forms, ten times each. In the second stage they fed it 420 high-frequency and medium-frequency verb pairs, 190 times each, of which 336 (80%) were regular. Frequencies were determined by published statistics of a large corpus of written English. The model responded accordingly: in the first stage, the regular pattern was only exemplified by two verbs, only one more than each of the eight irregular patterns, and the model in effect recorded ten separate patterns of associations between stem and past. Thus it performed perfectly on irregular verbs. In the second stage, there was a huge amount of evidence for the regular pattern, which swamped the associations specific to the irregular

verbs, resulting in 'overgeneralization errors' such as 'breaked'. Finally, as the 420 word corpus was presented over and over, the model was able to strengthen connections between features unique to irregular stems and features unique to its past forms, and to inhibit connections to the features of the regular ending, so as to approach correct performance.

The prediction, then, is that children's overgeneralization should also be triggered by changes in the ratio of irregular to regular forms in their vocabularies. The prediction is completely false. Fig. 1 shows data from four children at six different stages of development; overgeneralization typically occurs in the stage marked III. The proportion of regular verbs in the children's vocabularies remains essentially unchanged throughout this period, and there is never a point at which regular verbs predominate[8]. The same is true of token frequencies, and of frequencies in parental speech[8,12]. The cause of the onset of overgeneralization is not a change in vocabulary statistics, but some endogenous change in the child's language mechanisms. This is also shown by the fact that across a sample of children, use of the regular pattern correlates with general measures of grammatical sophistication, though not with chronological age. The use of irregular past forms, in contrast, correlates with chronological age[13]. This is exactly what one would expect if, contrary to the predictions of the model, rote (for the irregulars) and rule (for the regulars) were distinct mechanisms, the former depending on sheer quantity of exposure to the language, the latter on mastery of the grammatical system in general.

A second interesting way in which the model appears to mimic children is in the late appearance of doubly marked errors such as 'ated'. The model becomes prone to such errors because of response blending: when the responses for 'ate' and 'eated' each attain a sufficient level of strength at the same time, the model has no way of keeping track of which segments belong to which target and blends them. The alternative hypothesis is that children misconstrue 'ate' as itself being a stem and mistakenly attach the regular ending to it – basically, they think there are two distinct English verbs, 'eat' and 'ate'[14,15]. The data favor this hypothesis. Unambiguous blends of irregular vowel changes and the regular ending (e.g. 'sepped' for 'sip') are extremely rare in children's speech[8]. However, errors involving a past form misconstrued as a stem are common: children often say 'ating', 'he ates', and 'to ate'[15]. Moreover, when children are simply asked to convert 'eat' to a past tense form in experiments, they virtually never say 'ated', showing that when children do say 'ated', it is because of something they do to an input consisting of 'ate', not an input consisting of 'eat'. Apparently children do not derive inflected forms by haphazardly assembling them out of bits and pieces associated with the stem; they largely respect the integrity of words and the systematic modifications that can be applied to them[16]. In sum, the mechanisms invoked by the Rumelhart–McClelland model to account for children's behavior – lack of distinct mechanisms for memorization and rules, sensitivity to input frequency, and response blending – are inconsistent with the data from developmental psycholinguistics.

Implications for neuroscience

The Rumelhart–McClelland model is an extremely important contribution to our understanding of human language mechanisms. For the same reason that its impressive performance at first seemed to vindicate associative networks lacking implemented rules, the empirical flaws revealed by closer scrutiny provide valuable lessons about the kinds of mechanisms that language – and probably many other aspects of cognition – requires.

(1) *Elements versus their positions.* Lashley's problem of serial order in behavior applies in full force to language, and it is not solved by invoking feature units that conflate a feature and its immediate context. Such units cannot encode certain words at all, and they cannot explain patterns of psychological similarity defined by a given feature appearing in different serial positions.

(2) *Variables.* A variable is a symbol that can stand for a group of individuals regardless of their individual properties; in arithmetic 'x + 1 > x' is true regardless of whether x is even, odd, prime, and so on. Languages use variables in many of their

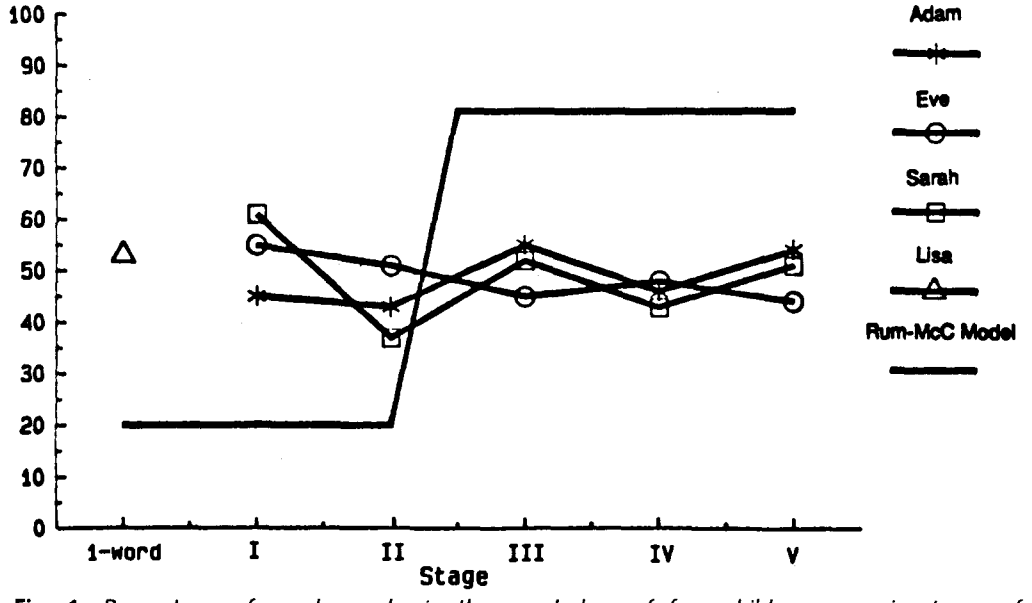


Fig. 1. *Percentage of regular verbs in the vocabulary of four children over six stages of development, and in the Rumelhart–McClelland model. The model predicts that 'overgeneralization errors' – i.e. giving an irregular verb a regular past-tense, like bring/bringed – would result from a high incidence of regular verbs in the vocabulary, which strengthens the connections for this 'regular pattern'. While the model showed overgeneralization errors after the introduction of a high percentage of regular verbs, there is no evidence that children use a high percentage of regular verbs at any stage that can be correlated with their overgeneralization errors.* (Taken with permission, from Ref. 8.)

operations; the regular past tense rule in English, which adds 'd' to the variable "stem", is a perfect example. Associating response features with the concrete features of a class of inputs is not the same thing as linking them to a symbol or variable that represents the class itself, because the associations are sensitive to the properties of the particular sample of inputs in a way that a genuine variable is not.

(3) *Individuals.* Two objects may share all their relevant features yet may be distinct in the world, hence their representations must be kept distinct in the brain. The case of 'lie/lay' versus 'lie/lied' shows that it must be possible to represent two identical patterns of features as corresponding to distinct entities. It is not enough to have large numbers of units with different perceptual receptive fields; some structures must be dedicated to representing an entity as simply being a distinct entity *per se.*

(4) *Binding.* Vision researchers have recently been made aware of one of the inherent problems of representing objects simply as patterns of activation over feature maps: it is impossible to keep the bindings of two simultaneously presented objects distinct, and a pressured perceiver is liable to illusory conjunctions whereby a red circle plus a green square is perceived as a green circle plus a red square[17]. A serial attentional mechanism is invoked in such cases to glue features into objects. Similarly, it is not sufficient that words be produced solely by activating patterns of features associated with an input, because when there are competing targets, there is no way of keeping the competing alternatives from blending. Simple connectionist models of language have this problem, but children's inflectional systems, apparently, do not.

(5) *Modularity.* It has recently become apparent that the visual system is not a single black box but is composed of many partially autonomous subsystems[18]. This conclusion was suggested by the methodology of 'dissection by psychophysics' even before it was corroborated by neuroanatomical and neurophysiological techniques. Though the neuroanatomy of language is not well understood, the equivalent psychophysical investigations strongly support a functional decomposition of language skill into subcomponents, and any model of language abilities will have to reflect this rather than mapping from input to output in a single link. Furthermore, the internal 'links' are organized in specific ways. In the present case, phonology and morphology reveal themselves as distinct subsystems, and that is only the most obvious cut.

(6) *Independence from correlational statistics of the input.* Connectionist networks, like all associationist models, learn by recording patterns of correlation among perceptual features. Language acquisition almost certainly does not work that way; in many cases, children ignore pervasive environmental correlations and make endogenously driven generalizations that are in some cases surprising with respect to the correlational statistics of the input but are consistent with subtle grammatical principles[19]. This can be seen in the case of the past tense, where the onset of overgeneralization is clearly independent of input statistics, and the extent of generalization in the adult state (e.g. avoiding it if an irregular form exists; but overriding the irregular form if the verb is derived from a noun root) is not a reflection of any simple correlational property of the input. More generally, to the extent that language is composed of separate subsystems, the role of environmentally driven changes must be quite circumscribed: if a subsystem's inputs and outputs are not connected to the environment, but to other internal subsystems, then they are invisible to the environment and there is no direct way for the connectionist's 'teaching inputs' to tune them to the correct state via incremental changes from a *tabula rasa.*

Overall, there is a more general lesson. Theories attempting to bridge neuroscience and cognition must be consistent with the data of both. The data of human language in particular are extremely rich, and theories of considerable sophistication and explanatory power have been developed in response to them. Though it may be convenient to impose a revisionist associationist theory on the phenomena of language, such a move is not scientifically defensible. Building the bridge will be more difficult, and more interesting, than it might first appear.

Selected references

1 Feldman, J. A. and Ballard, D. H. (1982) *Cognit. Sci.* 6, 205–254
2 Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. (1986) in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1: *Foundations*) Rumelhart, D. E., McClelland, J. L. and the PDP Research Group, eds), pp. 77–109, MIT Press
3 Crick, F. and Asanuma, C. (1986) in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2: *Psychological and Biological Models*) McClelland, J. L., Rumelhart, D. E. and the PDP Research Group, eds), pp. 333–371, MIT Press
4 Sejnowski, T. (1987) *Trends Neurosci.* 10, 304–305
5 Rumelhart, D. E. and McClelland, J. L. (1986) in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2: *Psychological and Biological Models*) (McClelland, J. L., Rumelhart, D. E. and the PDP Research Group eds,), pp. 216–271, MIT Press
6 Berko, J. (1958) *Word* 14, 150–177
7 Chomsky, N. (1959) *Language* 3, 26–58
8 Pinker, S. and Prince, A. (1988) *Cognition* 28, 73–193
9 Lachter, J. and Bever, T. G. (1988) *Cognition* 28, 195–247
10 Fodor, J. A. and Pylyshyn, Z. (1988) *Cognition* 28, 3–71
11 McClelland, J. L. and Rumelhart, D. E. (1985) *J. Exp. Psychol. Gen.* 114, 159–188
12 Slobin, D. I. (1971) in *The Ontogenesis of Grammar: A Theoretical Symposium* (Slobin, D. I., ed.), pp. 215–223, Academic Press
13 Kuczaj, S. A. (1977) *J. Verbal Learning Verbal Behav.* 16, 589–600
14 Kuczaj, S. A. (1978) *Child Dev.* 49, 319–326
15 Kuczaj, S. A. (1981) *J. Child Lang.* 8, 485–487
16 Slobin, D. I. (1985) in *The Crosslinguistic Study of Language Acquisition* (Vol. II *Theoretical Issues)* (Slobin, D. I. ed.), pp. 1157–1249. Erlbaum Associates
17 Treisman, A. and Schmidt, H. (1982) *Cog. Psych.* 14, 107–141
18 Van Essen, D. C. and Maunsell, J. (1983) *Trends Neurosci.* 6, 370–375
19 Pinker, S. (1984) *Language Learnability and Language Development*, Harvard University Press
20 Wickelgren, W. A. (1969) *Psychol. Rev.* 76. 1–15
21 Rosenblatt, F. (1962) *Principles of Neurodynamics*, Spartan