Kim Plunkett

Connectionists Approaches to Language Processing and Acquisition

Abstract

The degree to which the behaviour of parallel distributed processing (PDP) models approximates children's acquisition of inflectional morphology has recently been highlighted in discussions of the applicability of PDP to the study of human cognition and language. In this paper, an attempt is made to examine many of the limitations of the Rumelhart and McClelland model and adopt an empirical, comparative approach to the analysis of learning (i.e., hit rate and error type) in two sets of simulations in which vocabulary structure (class size and token frequency) and the presence/absence of phonological subregularities are manipulated. A 3-layer back propagation network is used to implement a pattern association task involving mappings which are analogous to the present and past tense forms of English verbs, i.e., arbitrary, identity, vowel change, and suffixation mappings. Several "competitions" in the learning of the four verb classes in networks which use random assignment of strings into the four verb classes are explored. The conditions under which different "default" transformations are employed and various overgeneralization errors appear (both "pure" and "blended" errors) given manipulations of the frequency of each mapping type (class size) and token frequencies (number of repetitions) are assessed. In a second set of simulations, an identical set of type and token frequencies are used, but strings in the identity and vowel change classes are assigned on the basis of phonological characteristics in the stem (e.g., identity stems end in a dental). These regularities are structured in ways that are analogous to English, e.g., characteristic but not predictive. Phonological cues are exploited by the system, leading to overall improved performance. However, overgeneralizations and competition effects continue to be observed in similar conditions. These simulations establish that characteristics of input frequency, in interaction with phonological subregularities, determine the types of errors produced by the network — including the conditions under which "rule-like" behavior will and will not emerge. The results are discussed with reference to behavioral data on children's acquisition of the past tense.

1. Introduction

The last decade has witnessed the emergence of a new approach to the study of human cognitive processes. The inspiration for this work has been a new computational architecture based on artificial neural networks. Neural networks modelled in computers share a number of properties in common with the human nervous system, including the brain. For example, they may contain many thousands of units or cells all highly interconnected with each other. Information is passed through the

network by the transmission of the excitatory or inhibitory signals emerging from the cells, along the pathways connecting the cells to each other. The state of the network is determined by the global pattern of activity of the cells, each cell influenced by the activity of its neighbours, and they in turn by their neighbours. Cells may have thresholds causing the sudden release of a signal when their energy levels exceed a certain value. Non-linearities permit these systems to take on complex characteristics which resemble the functioning of intelligent systems. From the point of view of human cognition, however, artificial neural networks possess a variety of properties beyond their apparent biological plausibility, which make them appealing for modelling purposes.

Artificial neural networks (ANNs) are inherently *parallel* processors. The connectivity of individual units in these computational architectures permit diverse sources of information to impinge upon the activity of any given node in the network. Human cognition, too, is massively parallel in its organisation, both across and within modalities. For example, it is known that on-line phonological processing is influenced by the syntactic categories of target words (Marslen-Wilson/Welsh 1978), and that linguistic processing can be influenced by perceptual information simultaneously available to the hearer (Palmer 1975). Information in an ANN is typically *distributed* throughout the system i.e. there is no single location where a given fact is represented. Rather representations are constituted by global patterns of activity in the network. This mode of representation leads to several crucial network properties. First, ANNs tend to be robust in the face of noisy input or damage to the connections in the network. Second, conceptual representations tend to have a *prototype* structure as opposed to a categorical structure. Recent work in cognitive psychology (Rosch 1973) attests to the 'fuzzy' nature of human concepts and it has long been documented (Lashley 1933) that damage to the central nervous system most often results in graceful degradation of performance rather sudden or complete loss of a skill.

ANNs are also able to learn. Given a particular task, they are able to self-organise their internal patterns of connectivity and activity to conform to a variety of mapping characteristics. In the past, the range of problems that could be solved were limited (Minsky/Papert 1988). Recently, powerful learning algorithms have been devised (Rumelhart/Hinton/Williams 1986) that enable ANNs to tackle a much wider range of learning tasks. Interestingly, learning in these systems is context sensitive. Changes to the network are determined by an interaction of the current state of

the network with the task demands from the environment. Thus, a network might be able to achieve major changes to its mapping characteristics when a task is relatively close to its current 'problem space' but make little progress on a task that requires substantially different forms of behaviour to that which the network is accustomed. In studies of human development, Piaget (1953) has pointed to the epigenetic nature of change in which the emerging human cognitive system is seen as an *equilibration* balancing the fundamental developmental processes of assimilation and accommodation. Similary, Vygotsky (1962) discusses the *Zone of Proximal Development* as a constraining factor on learning.

Despite, or perhaps by virtue of, the range of properties claimed for ANNs as appropriate computational architectures for modelling the fundamental aspects of human cognition, many cognitive scientists (Pinker/Mehler 1988) have argued that these parallel distributed processing (PDP) systems are unable to capture a number of essential characteristics of cognition. Hence, they are fatally flawed as plausible models or are seen not to offer any new insights into human *cognitive* functioning[1]. Most criticisms build upon the importance of viewing human cognition as a *symbolic* system. Cognitive representations are essentially symbolic in nature and cognitive processing consists in the manipulation or transformation of symbolic structures by sets of *rules* or *principles*, themselves couched in terms of abstract symbolic definitions. ANNs perform exclusively numerical calculations and though they may be capable of extracting subtle statistical regularities from complex data sets, they provide no principled account of the symbolic level of human cognitive functioning. A good deal of research in cognitive science during the past three decades has focused on language as a paradigm example of the symbolic mode of processing in the human cognitive system. The compositionality and systematicity (Fodor/Pylyshyn 1988) of linguistic structure are elegantly captured by the classical symbolic position. Processing models of comprehension and production have experienced considerable success within the confines of a symbol manipulating, serial von Neumann machine. In contrast, ANNs are internally unstructured, opaque to analysis and hence resistant to providing principled and predictive accounts of linguistic structure and processing. Similarly, it has been argued (Chomsky 1980) that the systematic nature of human language can-

[1] Though it is often admitted that ANNs may contribute to our understanding of the physiological underpinnings of cognition.

not be based on abstractions from the linguistic input that children hear during the course of acquisition. The 'poverty of the stimulus' argument requires that the human infant be endowed with some language acquisition device (LAD) which ensures that s/he will discover the correct grammar for his/her language in finite time and from an infinite range of possibilities. In contrast, PDP systems suggest that generalised learning algorithms are able to extract representations from impoverished linguistic input, adequate to the task of supporting an adult level of linguistic performance. It is no wonder that language has been the focus of attention in the paradigm conflict between connectionist and symbolic approaches to cognitive science.

In this article, we will review the current status of connectionist approaches to language processing and acquisition from the point of view of a single connectionist model. We begin by providing a brief overview of the operation of a simple ANN — the multi-layered perceptron. The Rumelhart/McClelland (1986) model of the acquisition of the English past tense is then described and evaluated. This model shows how a single, generalised learning mechanism can capture both the regularities and irregularities in the English past tense system without the need to postulate or endow the system with explicit rules. Certain inadequacies of the model are then reviewed and an alternative connectionist model of the acquisition of the English past tense is proposed. Although this combined work does not constitute a refutation of the classical symbolic approach to human cognition, it demonstrates that the connectionist perspective on language processing and acquisition is capable of providing principled accounts of cognition and has yet to be shown to be fatally flawed.

2. Representation and learning in an ANN

The basic features of the architecture of a multi-layered perceptron are illustrated in Figure 1.

fig 1

Figure 1: A simple multi-layered perceptron

The circles (*a–e*) represent the units or cells in the network whilst the uni-directional arrows depict the connections or pathways between the cells. In this particular network, the cells are organised in layers such that cells on the left (the input layer — *a, b*) feed into the cells in the middle (the hidden layer — *c, d*), which themselves feed into the cell on the right (the output layer — *e*). There are no intra-level connections. Activity in the network is initiated by the presentation of signals at the input layer. For example, an input of value "0" might be presented to cell *a*, whilst an input of value "1" might be presented to cell *b*. Cells *a* and *b* respond to the input signals by becoming activated and remaining dormant, respectively. In this case, we have arranged that the input cells have *linear activation functions* i.e., the cells simply take on activation values corresponding to the values of the signals to which they are exposed — 1 and 0.

Next, the activations of cells *a* and *b* are propagated to the next layer in the network, the hidden units. Each unit in the input layer is connected to both units in the hidden layer. Thus, unit *c* receives stimulation from both unit *a* and unit *b* — likewise, unit *d*. However, rather than stimulating the hidden units directly, the activation on the input units are modulated by the strength or *weight* of the connections leading to the hidden units. The weight of a connection may vary from being large and positive to large and negative. In the former case, we consider the connection between two units to be strongly excitatory. In the latter case, the connection is strongly inhibitory. In the network in Figure 1, the connections leading to both the hidden units consist of a single inhibitory and a single excitatory weight. Since the weights of the connections modulate the activation values of the input units, then the net effect of an input unit on any given hidden unit is determined by the *product* of the unit's activation value and the value of the connection leading to the hidden unit.

Suppose that the activation value of unit *a* is a_a and that the value of the connection leading to unit *c* from unit *a* is w_{ca} then the net effect of unit *a* on unit *c* is $a_a w_{ca}$. Similarly, the net effect of unit *b* on unit *c* is $a_b w_{cb}$. Thus, the total net input to unit *c* from the input layer is determined by the summation $net_c = a_a w_{ca} + a_b w_{cb}$. Substituting the values of the connections indicated in Figure 1 and the activation values on units *a* and *b*, we obtain the expression $net_c = 1.0 \times 1.0 + 0.0 \times -1.0$. In other words, the net input to unit is 1.0. Similarly, the net input net_d to unit *d* is given by the expression $net_d = a_a w_{da} + a_b w_{db}$ which yields a value of –1.0. In general, the net input net_i to a unit *i* is determined by the sum of

the weighted activations feeding into that unit. If there are j such weights, then the expression

eq 1

computes the net input to unit i, where a_j is the activity on unit j and w_{ij} is the value of the weight connecting unit j to unit i.

Unlike the units in the input layer the units in the hidden layer are *linear threshold units*, i.e., unless the input to a linear threshold unit reaches some threshold value θ, then the unit does not become active. In Figure 1, the threshold is set such that $\theta = 1.0$. Futhermore, the response properties of the units are set such that if input to a unit exceeds the threshold value, the output from the unit remains constant. The response property of the linear threshold unit is, thus, a *step function* (see Figure 3, page 41). If θ is the threshold value of the linear threshold unit i and net_i is the net input to unit i, then the response o_i is determined by the formula:

eq 2

We have determined that $net_c = 1.0$ and that $net_d = -1.0$. Given that units c and d have the activation function described in Equation 2, then the outputs from c and d will be 1.0 and 0.0, respectively. Propagating these signals from the hidden layer to the output layer (unit e which is also a linear threshold unit), the reader can determine that the output from the network will be 1.0. In fact, the reader can determine the mapping contingencies for a variety of input patterns and observe that the ANN in Figure 1 maps the boolean function EXCLUSIVE OR (see Table 1).

ta 1

The above example illustrates that the mapping properties of a network are determined entirely by the matrix of weighted connections in

the network and the activation functions of the units in the network. We may interpret this configuration as the network's representation of the mapping problem. Thus, changes to the values of the weight and activation parameters will, in general, alter the mapping properties of the network i.e. change its representation. However, there are several other properties of the network in Figure 1 that are worth highlighting here. First, the network processes the input in *parallel*. Unit *c* takes into account information from both units *a* and *b*. Similarly, unit *e* processes units *c* and *d* in parallel. Second, information about the mapping relations is *distributed* throughout the network. For example, it is not possible to point at a single connection or unit in the network which 'represents', say, the mapping relation $10 \rightarrow 1$. The whole nexus of connections represents this property. But the same nexus of connections represents the other mapping relations of EXCLUSIVE OR (Table 1). By using distributed representations, the network is able to *superimpose* a variety of mapping relations upon a single set of connections.

The distributed character of the representations in an ANN tend to make them robust to damage or noisy input. Although the mapping characteristics of the network illustrated in Figure 1 will change if the connections or input is changed, in larger networks performance tends to *degrade gracefully*, rather than alter categorically. Since information about any single mapping relation is spread over a wide range of connections and units, then damage to any one (or few) of them will have little effect on performance. Performance will tend[2] to deteriorate gradually as the number and/or size of the distortions increase.

The network in Figure 1 has been hand-wired, i.e., we have ourselves selected a set of connections that will perform the desired mapping functions. However, it is possible to *train* a network of this kind to set its own connections, starting from some random configuration, such that the appropriate mappings are achieved. The training procedure consists in gradually adjusting the current set of connection weights in response to an error signal from the external environment. The network in Figure1 needs to be supplemented with two components in order for training to occur. First, the network requires some mechanism to discover the "error" that it produces in response to any given input. This is achieved by calculating the discrepancy between the desired output and the actual

---

2 The hedge *tend to* is preferred since ANNs *can* also behave in a categorical fashion under certain conditions. This is easily observed by altering the weights in Figure 1 and computing the resultant output. However, similar phenomena may occur in larger networks.

output from the network. We will call this the *error signal*. Second, the network requires a *learning algorithm* that is able to translate the error signal into a set of weight changes that reduces the error produced in response to a given input. It is important to note that the learning algorithm is ignorant of the final solution to the problem — otherwise, it could simply set the appropriate weights immediately, just as we did in Figure 1. Rather, the learning algorithm examines whether changing a weight in a particular direction is likely to contribute to a decrease in the error. It then adjusts the weight by a small amount in the appropriate direction. Changes to one weight in the network are made without regard to weight changes in other parts of the network. We say that weight changes are based on *local* computations. When the learning algorithm has made changes to all the weights in the network, we can test to see if the error has been reduced or we can continue to train the network on the same or other input/output pairings.

A useful way to conceptualise the process of minimising the error is to consider the manner in which a single weight value can effect the overall error for a given input pattern. Figure 2 plots a hypothetical function relating the value of a single weight in a network to the error it would produce on the output if all other weights were left constant.

Figure 2: An hypothetical error/weight space

Points A, B and C correspond to three different weight values. If the network is in a state represented by point A, then the learning algorithm will attempt to increase the value of the weight. If the network is in a state represented by C, it will attempt to decrease the value of the weight. In essence, the learning algorithm evaluates the *slope* of the error function. If the slope is positive the weight is decreased. If the weight is negative

then the weight is increased. Equation 3 summarises this relation

eq 3

where Δw_{ij} is the change in the weight connecting unit j to unit i and E_p is the error produced at the output when the network is presented with pattern p. The learning algorithm is often referred to as a *gradient descent* algorithm. Since the algorithm performs calculations based entirely on *local* information (i.e., the slope of the error function), it cannot tell whether the weight changes are moving towards a local minimum in the error function (as in A) or a global minimum in the error function (as in C). In particular, if the network is in a state represented by the point B, where the error gradient is negative in one direction and positive in the other direction, the network has no means of deciding upon the most appropriate direction to move in weight space. If the network starts off in the state A, it is likely that it will reach the trough of the local minimum at which the slope of the error curve i.e. $\partial E_p/\partial w_{ij}$, in Equation 3, is zero. In this circumstance, the term Δw_{ij} in Equation 3 will be zero and the weight will never be changed i.e., the network will be stuck in a local minimum where the input is mapped erroneously.

Although local minima are sources of permanent error in network training, there are ways of avoiding them. First, the weight represented in Figure 2 is not the only weight in the network. Changes in the other parts of the network will change the error function depicted. This itself may cause the local minimum to move or disappear[3]. Second, the network may be able to "hop over" that part of the weight space containing the local minimum. For example, if the learning algorithm forces a large change to the weight in Figure 2 when it is at state A, then the change may be adequate to move it directly to a point to the right of B. The amount by which a weight is adjusted on any given learning trial is determined, amongst other things, by a scaling parameter called the *learning* rate. Learning rate can be thought of as a constant of proportionality in Equation 3. If learning rate is large, then the value of Δw_{ij} in Equation 3 will be large.

In order for a learning algorithm to minimise the error in the output by performing gradient descent, then the partial derivative of the error

---

[3] Though note that it may also cause the global minimum to move as well.

function, i.e. $\partial E_p/\partial w_{ij}$, in Equation 3, must be definable. In other words, there must be a way to calculate the slope of the error function. The term E_p itself is derived from the discrepancy between the desired output from the network and the actual output of the network. If t_p is the desired output and o_p is the actual output, then

eq 4

However, the value of o_p is determined by the activation function of the output unit. Assuming that the activation value of a unit is the same as the output from the unit, then this yields

eq 5

where a_p is the activation value of the output unit when the network is presented with pattern p. Substituting this in Equation 3 we get

eq 6

Since the term t_p is a constant, the derivative $\partial t_p/\partial w_{ij}$ is zero. Hence

eq 7

In the network illustrated in Figure 1, the activation function of units *c*, *d* and *e* is the step function (Equation 2). Unfortunately, we cannot determine a derivative for the step function as is required by Equation 7. Thus, we cannot train a network that is made up of linear threshold units with a learning algorithm designed to perform gradient descent. Instead of using linear threshold units, it is common to use units which have a logistic activation function as in Equation 8.

eq 8

where net_i is the net input to unit *i*. The logistic function is differentiable. The difference between the step function and the logistic function (sometimes known as the squashing function or sigmoid function) is shown in Figure 3.

fig 3

Figure 3: Activation functions

The logistic function has a definable slope at all points whilst the step function has a discontinuity at 1.0.

The most commonly used learning algorithm based on gradient descent is *Back Propagation* (Rumelhart/Hinton/Williams 1986). Back propagation uses the error signal on the output to adjust the weights connecting the hidden units to the output units. "Error" is also assigned to the hidden units so that changes can be made to the weights connecting the input units to the hidden units. Error is assigned to the hidden units by taking a weighted sum of the error on the output units. The algorithm assumes that if a hidden unit is highly active when a particular pattern is presented to the network, then the highly active hidden unit is apportioned a substantial portion of the blame for the output error. Thus, error is propagated backwards through the network in response to the output error.

A multi-layered perceptron supplemented with a back propagation learning algorithm is, in principle, capable of solving any mapping problem that can be construed as a smooth mathematical function[4] (Cybenko 1989). Typically, the network solves the mapping problem by searching for predictive regularities in the set of input patterns and constructing representations of these regularities at the hidden unit level. Hidden unit representations may be distributed or localised to a single unit. For example, a single hidden unit may take on the task of detecting recurring features in the input set and exploit the predictive value of these features in constructing appropriate mapping characteristics. The features thereby "discovered" by the network need not be limited to local properties of the

[4] Though as we have noted above, the network may get trapped in a local minimum.

input patterns, but may themselves be complex distributed features that may not be obvious to the casual human observer. It is important to realise that the representations constructed by the network during the course of learning are "contextually embedded". That is, the configuration of weights in the network can be said to *represent* a solution to a given problem, but only in interaction with the environmental domain. It is more appropriate to interpret the knowledge that is stored in a network's weight matrix as a *medium* for expressing a set of relationships rather than as a *model* of the environment to which it is exposed. The weights in a network have no meaning in and of themselves. They are, indeed, just numbers. However, in interaction with an appropriate environment, the weights take on the capacity to express certain facts about that environment.

3. Learning the English past tense

It is a common finding in both naturalistic and experimental contexts that English speaking children sometimes produce erroneous past tense forms, such as *goed* or *sitted*, in which /-ed/ is added to verb stems whose past tense forms are exceptions to the regular rule (Bowerman 1982, Bybee/Slobin 1982, Derwing/Baker 1986, Kuczaj 1977, Marchman 1984). The occurrence of these errors is typically thought to illustrate that children are capable of going beyond their data to create novel lexical forms which they are not likely to hear in the input. Interestingly, overgeneralisations typically occur *after* children have been using correct forms of irregular verbs appropriately. With development, the organisation of the linguistic system supports the correct production of both regular and irregular past tense forms. This apparent regression and subsequent improvement suggests that acquisition involves a stage-like reorganisation of rules and representations (Bowerman 1982, Karmiloff-Smith 1979, Karmiloff-Smith 1986, Pinker/Prince 1988) and is an oft-cited example of U-shaped development (Bever 1982, Strauss 1982). Taken together, the phenomena of overgeneralisations and U-shaped acquisition have been viewed as among the most persuasive pieces of behavioural evidence that language learning involves the process of organising linguistic knowledge into a system in which rules and the exceptions to those rules must coexist.

Acquisitionists have not generally questioned whether children use rules in learning and producing language. Indeed, it would appear to be difficult to account for many phenomena of acquisition, most notably overgeneralizations, without some version of a rule system. Debate has

instead focussed on what rules are acquired, what form they must take, how and when children *do not* appear to utilise an adequate version of the rule system, as well as how and when the correct version is eventually attained. In addressing these questions, it is assumed that the input itself does not force the child to begin to produce overgeneralisations, nor to eliminate those errors from their output. Rather, endogenous factors trigger reorganisational processes that result initially in a performance decrement followed by gradual mastery of the system.

Recently, work within the connectionist perspective has promoted a reevaluation of several of the basic assumptions about the constructs and processes guiding the acquisition of language. In an attempt to illustrate the applicability of parallel distributed systems to the "favored domain of non-associationist, higher-order structural cognition" (Maratsos 1988, p. 242), Rumelhart/McClelland (1986) set out to capture several of the facts of the acquisition of the English past tense. In general, the goal of this work was to suggest how a model of language processing and acquisition might be able to avoid reliance on rule-based mechanisms and discrete symbols, yet still capture what children do at various points in acquisition. Models such as this one characteristically utilize distributed representations and focus on elaborating the microstructure or sub-symbolic nature of cognition and language (Smolensky 1988b).

Rumelhart and McClelland's past tense simulation contains three major components (see Figure 4). First, an encoding device takes the present tense stems of English verbs, symbolised as binary characters, and converts each stem into a distributed representation of context sensitive phonological features called Wickelfeatures. Output from the encoding device consists of a vector of activation across the set of output units (460 in all)[5]. Secondly, a single-layered, pattern association network maps the set of Wicklefeatures, which it takes as input, onto a set of output units. The activity on these output units constitute the Wickelfeature representation of the past tense form of the verb that was originally presented to the simulator in its present tense form. The task of the pattern association network is to learn to map correctly input to output vectors through adjusting the set of weights which connect the input and output units. The adjustment of the weights is achieved by using the error signal obtained from comparing the actual output of the network with the de-

---

[5] Since the details of the encoding process are not of direct concern for the present article, the reader is referred to the original source for further information (see also Pinker/Prince (1988) and Bever (1989) for reviews and criticisms.

sired output stipulated by a teacher signal. The weights connecting the input and output units of the network are then adjusted using the Perceptron Learning rule[6]. This second component of the simulator is entirely responsible for the learning that is required to map present tense stems of verbs onto their corresponding past tense forms. This mechanism, then, can be seen to be the foundation for the overgeneralisations reported by Rumelhart and McClelland. The third component of the simulator takes as its input the vector representing the activity of the output units of the pattern associator. Its function is to generate the set of Wickelphones that best fit the output vector description. In principle, the decoder provides a Wickelphone description of the past tense form of the verb that was originally provided in the Wickelphone representation of the present tense stem to the encoder. Several researchers as well as Rumelhart and McClelland themselves have acknowledged several difficulties with this type of decoding process (Pinker/Prince 1988). The usefulness of Wickelfeatures as a technique for encoding linguistic information in networks of this type is not crucial for the issues discussed in this paper, and the reader is referred to the original source for details.

fig 4

Figure 4: Network architecture and verb performance in the Rumelhart/McClelland (1986) simulation.

The performance of the Rumelhart and McClelland simulation is important because the learning curves and overgeneralisations generated by the simulation resemble many of the errors and stages of development that children make and pass through in the acquisition of past tense verb forms. Figure 4 shows the “U-shaped” dip for irregular verbs during the

---

[6] The perceptron learning algorithm is a restricted version of the back propagation learning algorithm described in Section 2. It is restricted in the sense that it cannot be used to adapt the weights of networks within hidden units.

early stages of learning. This regression represents the stage of learning in which irregular verbs are treated as though they are regulars. Even more impressively, Rumelhart and McClelland's simulation provides distinct learning curves for the different classes of irregulars which closely map the types and relative timing of errors made by young children. For example, Kuczaj (1977) reports that past tense errors of the form "ated" occur later in development than errors which simply "add-ed" to the present tense stem ("eated"). Rumelhart and McClelland's simulator is also delayed in producing these former types of error.

More controversially, the Rumelhart and McClelland model (and the general class of models that it represents) does not rely in any obvious way on rules which are "assumed to be an essential part of the explanation of the past tense formation process" (Pinker/Prince 1988, p. 79). As Rumelhart and McClelland claim, "we have shown 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" (Rumelhart/McClelland 1987, p. 246). The ability of networks of this sort to mimic children's behavior when learning the past tense is intended to challenge the traditional view that acquisition is *necessarily* a process of organising and reorganising explicitly represented rules and principles, and their exceptions. These proposals have been met with enthusiasm in some circles, fueling many explorations of PDP models in other linguistic and non-linguistic domains (Churchland/Sejnowski 1988, Elman 1988, Elman/Zipser 1988, Hare/Corina/Cottrell 1989, MacWhinney et al. 1989, Mozer 1988, Seidenberg/McClelland 1989, Smolensky 1988a). Elsewhere, these claims have undergone considerable scrutiny and have met with resistance (Pinker/Mehler 1988). Several criticisms specifically address the details of the structure and/or success of this particular simulation. Others have been offered at a more general level, nominating it as the test case for evaluating the general potential of connectionist approaches (Fodor/Pylyshyn 1988).

Clearly, the Rumelhart and McClelland simulation has several substantive limitations as a model of children's morphological acquisition. The task modeled by this simulation cannot be said to resemble the task of language learning in any real sense. It is clear that children do not hear stem and past tense forms side-by-side in the input in the absence of semantic information or outside of a larger communicative frame. Nor do children receive an explicit teacher signal as feedback about the relationship between the phonological form of their output and what the correct

form should be. However, it is possible to characterise the Rumelhart and McClelland simulation at a more abstract level, as modelling an hypothetical, internal system-building process, such as *primary explicitation* outlined by (Karmiloff-Smith 1986). Other criticisms have focussed on the limitations of the phonological notation and the encoding/decoding processes used by Rumelhart and McClelland. For example, Lachter/Bever (1988) point out that Wickelfeature representations presuppose a theory of the phonological regularities present in the English past tense system. Lachter and Bever accuse Rumelhart and McClelland of using several "TRICS" (The Representations that It Crucially Supposes) in order to ensure that the model is sensitive to the linguistic properties of past tense formation and hence, performs in the way that it does.

More importantly for our purposes, these reviews point out that Rumelhart and McClelland misrepresent the input set within which children abstract and organize the regularities of the past tense system in three crucial ways. First, in the Rumelhart and McClelland simulation, one token each of the 18 most frequent verbs in English (16 of which happen to be irregular) is presented to the simulation during the first 10 training epochs. At that point in the learning process, the size of the input set is increased so that it is composed of a larger vocabulary of both frequent and infrequent verb forms. Pinker/Prince (1988) point out that the simulation's U-shaped developmental curve is likely to be a direct result of the *discontinuity* in vocabulary size and structure to which the network is exposed. It is no accident that the simulation's overusage of the /-ed/ ending and the related drop in performance on the irregular verbs coincides directly with the increase of the number of regular verbs in the vocabulary. While this vocabulary configuration does capture certain characteristics of the input to which children are exposed, generally accepted learnability conditions suggest it unwise to develop a model of acquisition which assumes that children receive a subset of the available linguistic data early in development.

Second, in the Rumelhart and McClelland model, exemplars (i.e., tokens) of particular verbs are presented with equal frequency. Bever (1989) suggests that Rumelhart and McClelland:

"predigested the input for their model in much the same way a linguist does — by ignoring real frequency information. This is probably the most important trick of all — and it is absolutely clear why they did it. Irregular past tense verbs are by far and away the most frequently occurring tokens. Hence, if Rumelhart and McClelland had presented their

model with data corresponding to the real frequency of occurrence of the verbs, the model would have learned all the irregulars, and might never receive enough relative data about regulars to learn them." (p. 11)

Third, Rumelhart and McClelland's failure to capture basic categorical differences between regular and irregular verbs is interpreted as a significant and fatal shortcoming of the model. According to Pinker and Prince, symbolic and PDP models share several assumptions about linguistic systems. Both classes of models are theoretically capable of dealing with type-frequency sensitivity, graded strength of representations, and competition among candidate hypotheses. However, the approach embodied in the Rumelhart and McClelland simulation differs from a rule-based one in its treatment of regular and irregular verbs:

1. Regular and irregular verbs are not distinguished qualitatively in terms of the phonological characteristics of individual members of a class or classes taken as a whole.
2. Phonological and morphological operations are applied uniformly to all verbs (in the formation of past tense forms) rather than differentially to regulars versus irregulars.

These distinctions are crucial components of Pinker and Prince's model of past tense acquisition. In their view, membership in the regular class is not dependent on phonological characteristics of the stem nor on the degree of phonological similarity among class members. The application of the regular rule occurs to verb stems regardless of phonological shape, and constitutes the default past tense formation procedure.

On the other hand, the stem and past tense forms of irregular (strong) verbs are stored independently in the lexicon. The past tense forms of irregular verbs are memorised as distinct lexical items, and are not derived from the stem. Further, most classes of strong verbs are characterised by family resemblances of phonological similarity, and are categorised as such with reference to lexical and morphological information. However, these phonological properties do not *guarantee* membership in a particular irregular class. Rather, the irregular verbs are

> "held together by phonologically unpredictable hypersimilarities which are neither necessary nor sufficient criteria for membership in the classes."
> (Pinker/Prince 1988, p. 122)

Thus, the acquisition and formation of regular and irregular past tense forms require two distinct mechanisms. However, the approach to past tense acquisition embodied in the Rumelhart and McClelland model

incorporates only one of them: the abstraction of family resemblance clusters of phonological similarity. According to Pinker and Prince, this mechanism can only do half of the job, as it is neither necessary nor appropriate for the acquisition of verbs in the regular class, since the operation of the regular rule is not sensitive to phonological regularity. Missing from the Rumelhart and McClelland model are higher-level lexical representations manipulated by the past tense rule regardless of their lower-level phonological character.

In this paper, I show how the Rumelhart/McClelland work can be enhanced, presenting two sets of simulations that explore learning in networks required to master mappings analogous to present and past tense forms in English. The simulations described in this paper differ from Rumelhart/McClelland's work in several respects. First, Wickelfeature representations are not used. Second, a three-layer back propagation network is adopted. Third, an empirical, comparative approach, systematically investigates the role of token frequency and the presence/absence of phonological subregularities on learning within networks of this type. Vocabulary structure is manipulated within a stable type frequency configuration, but the number of repetitions of each unique token that the network "sees" on each training epoch is varied. In the first set of simulations, membership in the classes of regular vs. irregular verbs *cannot* be determined by phonological information. However, in the second, several "TRICS" are exploited by specifically constructing vocabularies in which class membership is predictable, in ways that are analogous to English, by the phonological characteristics of stems. Lastly, no discontinuities are introduced into the learning set in any simulation.

4. Method

The simulations use an artificial language[7] that consists of randomly generated, legal (i.e., possible) English CVC, VCC and CCV strings. Each consonant and vowel is represented by a pattern of features distributed across 6 units, reflecting phonological contrasts such as voiced/unvoiced, front/middle/back, etc. The suffix representation (2 units) is not phonological. However, the network can use features of the final phoneme in the stem to decide which suffix units should be activated. The suffix units represent the allomorphs of the past tense morpheme in

[7] For a more complete description of the methodology (vocabulary and phonological representation), the reader is referred to Plunkett/Marchman (1989).

English, e.g., /-t/ following voiceless stop. Twenty units are used to encode each stem and past tense form.

Approximately 32 vowel transformations occur in English, e.g., /i/ ⇒ /A/, ring ⇒ rang; /u/ ⇒ /A/, come ⇒ came. In this language, a representative subset of 11 vowel transformations were chosen. Not all vowel change transformations are absolute, i.e., a vowel can be transformed to one, two or three possible new vowels in the output.

In all simulations, the network learns four types of mapping. Thus, the network, like the child, must learn to deal with several different classes of transformations simultaneously. However, in the *parent* simulations, the network is at a slight disadvantage in that strings are assigned to the different classes *randomly*, i.e., there is no more phonological similarity between the members of a given class than between members of different classes. The only exception is the vowel change class in which class assignment is conditional upon the possession of a vowel which can undergo a legal transformation. In the *phone* simulations, in contrast, we partially mimic the phonological subregularities which characterize the vowel change and identity verbs in English.

The members of the 4 classes are assembled from a "language" of 700 legal strings. The number of strings in each class (type frequency) is varied across simulation. The number of repetitions of a unique string (token frequency) is also manipulated so that the network experiences some items more frequently than others within a given sweep through the data. However, the total vocabulary is held constant (500 unique strings). All simulations were run on the "rlearn" simulator (Center for Research in Language, UCSD) using a back propagation learning algorithm, and contained 20 input units, 20 output units, and a hidden layer of 20 units. A disadvantage of this architecture is that the model is restricted to processing fixed length strings.

Performance is assessed in terms of the percentage of correct outputs in each class of stems. For incorrect outputs, categories of errors are tabulated, i.e., consonant miss(es), a vowel miss or a suffix miss. The closest phonological representation is also computed for each output, in order to estimate the actual "verbal" output of the network, and generate categories of error types (see tables 2 and 3).

5. Results and Discussion

5.1 Parent Simulations

Previous results from several series of simulations using this architecture (Plunkett/Marchman, 1989) showed that type and token frequency significantly affects learning in networks of this type. Class size and frequency of exemplars affect both the rate of learning and the final level of performance within that class. In addition, these parameters affect the degree to which characteristics of the mapping in one class will be adopted by the network when forming the past tense forms of verbs in other classes. In general, variations in token frequency appear to have a *greater* effect than type frequency. However, the effects can be observed in many directions, depending on which strategy is dominant in that simulation. Dominance of a particular strategy is determined by the relative type and token frequencies of the competing classes, in interaction with the global characteristics of the total mapping function that the network is required to perform. A noteworthy characteristic of these networks is their inability to map many arbitrary stems simultaneously.

These "network facts" are informative for a model of children's acquisition of language in only a limited sense, i.e., to the degree that the particular input configuration used accurately represents input to children. However, it is extremely difficult to determine the relative numbers of verbs of each type that are relevant and/or salient for a child. In the *parent* simulations, we settle on a representative configuration of the relative type frequency distributions in English in constructing our class sizes (Plunkett/Marchman, 1989) and then vary token frequency parametrically across simulations in an attempt to achieve optimal learning in all four verb classes. Table 2 outlines the type and token frequencies for the different mapping classes in the *parent* simulations.

5.3 Results of the Parent Simulations

Table 3 presents the overall hit rate % (after 50 epochs) and several different categories of errors for all of the subsequent *parent* simulations[8]. Errors on the arbitrary mappings are not presented due to their generally high level of performance. All simulations have been replicated with different string assignments to the various classes.

The results of these simulations will only be summarized briefly here

---

[8] Table 4 presents the same categories of errors for the *phone* simulations.

tab 2

Table 2: Parent simulations: Type (Class Size) and
Token Frequency distributions

(see Plunkett/Marchman, 1989). In general, type and token frequencies play an important role in determining the performance of a given class and the extent to which overgeneralization errors were observed. When type frequency (class size) of an irregular mapping is low, increasing the token frequency of that class results in a high level of performance for that class without any deleterious effects on the dominant form (highest type frequency) of mapping. However, if the type frequency of the irregular class is relatively large and is backed up by a high token frequency, then the performance in the dominant form of mapping deteriorates dramatically. In the case of the arbitrary mappings, successful learning occurs when the class contains a low number of tokens (class size is small) while each exemplar is presented to the system fairly frequently (token frequency is greater than 20). Because of the initial biases of these networks to perform an identity mapping process, performance on the arbitraries is poor unless these type/token constraints are met. When the input provides enough exemplars so that the arbitrary mappings can be mastered, they are generally unaffected by, and do not affect, other mappings in the network.

tab 3 (horisontal) full page

The error categories presented in Table 3 and the corresponding "Phone" table are to be interpreted as follows[9]:

Regular Errors

Inap Suf	The Stem is suffixized but with the wrong suffix.
Iden	The stem is treated as an identity stem.
Inap Vow-S	The stem is appropriately suffixized but undergoes an illegal vowel transformation.
Bld	The stem is appropriately suffixized but undergoes a legal vowel change transformation.
Vow Chan	The stem is treated as though it were a vowel change stem. description

Identity Errors

Suf	The stem is treated as a regular stem.
Inap Suf	The stem is treated as a regular stem but inappropriately suffixized.
Vow Chan	The stem is treated as though it were a vowel change stem.
Inap Vow	The vowel change transformation is "illegal".

Vowel Change Errors

Suf	The stem is treated as a regular stem.
Iden	The stem is treated as an identity stem.
Inap Vow-S	The stem is transformed as though it were a vowel change but the vowel change is "illegal". Furthermore, the stem is appropriately suffixized.
Inap Suf	The stem is treated as though it were a regular but inappropriately suffixized.
Bld	The stem undergoes a legal vowel change but is also appropriately suffixized.
Inap Vow	The vowel change transformation is inappropriate though may or may not be "illegal".

When natural languages incorporate arbitrary forms, they are generally highly frequent and constitute a relatively small class of items. For the young child acquiring language, these typological characteristics undoubtedly contribute to the early learning of these forms. However, children often overgeneralize the regular mappings to the arbitrary class (*go* ⇒ *goed*), but eventually learn the correct form. We also observe this effect in many of the *parent* simulations. However, unlike the Rumelhart/Mc-

---

[9] All figures are percentages of total errors after 50 epochs. Failure to sum to 100 in a few simulations indicates that a small percentage of other types of errors were also found. The hit rate gives the percent overall word hit rate after 50 epochs for each simulation. Errors on the arbitrary mappings are not presented due to their generally high level of performance.

Clelland simulation, this result cannot be a result of introducing a discontinuity in the vocabulary to which the network is exposed. In these networks, overgeneralizations on arbitrary mappings arise from the need to satisfy a variety of constraints *within the framework of a single mechanism*. The network is forced to reorganize its weight matrix to meet the requirements of the dominant form of mapping. Once this is achieved, however, the network reestablishes correct performance so that arbitrary forms may peacefully co-exist with stems from the other classes.

In addition, we observe that the regular and vowel change mappings frequently compete with each other for network resources in such a manner that neither class can be completely mastered simultaneously. Competition effects result in the production of many types of overgeneralization errors, in some situations, leading to complex patterns of "leakage" of mapping characteristics across classes. For example, vowel change overgeneralizations to the regular class may occur at the same time as suffixation overgeneralizations occur to the identity class. Blended errors are also observed (i.e., the application of two mapping regularities in a single form). In studies of children's acquisition of the past tense in English, sub-regularities in the irregular system sometimes give rise to their own patterns of overgeneralization, albeit less frequently than the standard "add -ed" overgeneralization (Bybee/Slobin, 1982). That is, children will sometimes "overgeneralize" a vowel change or identity mapping to a regular or irregular stem, producing errors such as *pick* ⇒ *packed*, or combine mapping types to produce blended responses, such as *ated*.

In the *parent* simulations, assignment of stems to each of the 4 classes was randomized, and thus, the phonological character of the input stems could not predict category membership. Yet, many examples of overgeneralizations were found, of both the standard and irregular variety, given certain type and token frequency configurations. However, *none* of the overgeneralization errors that were observed in these simulations can be attributed to the phonological structure of the input set. Finally, none of these simulations succeeded in reaching "adult-like competence" in all mapping classes simultaneously. The *phone* simulations explores whether the addition of phonological predictability into the input set will enable this system to master the past tense.

5.3 Phonological Simulations

English irregular verbs possess phonological properties that hold

together the members of a given class. Strings which undergo an identity mapping end in a dental consonant (e.g., *hit*), and vowel change verbs exhibit family resemblance clusterings (e.g., *ring* ⇒ *rang* ; *sing* ⇒ *sang*). Although these properties characterize the classes, they are nevertheless insufficient to reliably predict membership in these irregular classes; e.g., many regular stems also end with a dental. As discussed by Pinker/Prince (1988), the phonological structure of the irregular classes is crucial to the hypothesis that different mechanisms of past tense formation operate on the irregular and regular verbs. The regular rule is applied generally, without reference to the properties of the stem; whereas, irregular transformations take this phonological information into account in the production of a past tense form. According to studies of acquisition, children sometimes do produce overgeneralizations of vowel changes or identity mapping in addition to the standard overgeneralization error (Marchman, 1988). But, because irregular forms are not formed using a "default" rule, errors of this type are not seen as true "overgeneralizations" in the same sense as the standard overapplication of the "-ed" suffix, but are instead thought to result via analogy to the phonological shape of individual stems (MacWhinney 1987, Pinker/Prince 1988).

In the *phone* simulations, we partially mimic the conditions of membership in the irregular verb classes of English, by imposing the following constraints on class assignment:

- All identity stems must end in a dental.
- All vowel change stems are restricted to eleven possible VC endings.

We also ensure that the regular class contains stems possible irregular stems, e.g., some regulars end in a dental. Two questions are relevant:

1. Can the network exploit the phonological sub-regularities in the identity and vowel change classes? i.e., do the additional constraints on class membership aid the discovery of class memberships leading to improved performance?
2. Do patterns of competition and overgeneralization occur when phonological sub-regularities are available to the network that are similar to those when phonological information is not available to the network?

We have conducted 19 *phone* simulations which repeat the type and token frequencies of corresponding *parent* simulations (see Table 2). The *phone* simulations represent a second approximation to the task facing

the young child learning the relationship between the present and past tense forms of the verbs of English.

5.4 Results of the Phonological Simulations

Here, we provide a summary comparison of the *phone* to the *parent* simulations. In general, all phonological simulations exhibit a higher level of performance, across mapping types, compared to the *parent* simulations, except for the arbitrary mappings in a few simulations (see Plunkett/Marchman, 1989). Note, however, that the regulars perform minimally better under the *phone* condition. The greatest differences tend to occur when the token frequency of the *vowel change* class is relatively high (simulations 5, 17, 18 and 27). Since there are no differences between conditions other than the sub-regularities in the identities and vowel changes, we can attribute the lower performance of the regulars in the parent condition to the absence of these subregularities. In the *phones*, the phonological subregularities conspire to protect the regulars from interference, despite the facts that (a) the regular class contains stems that resemble the vowel change and identity classes (similar vowel and final consonant), and (b) there are no explicit features marking the regular stems as "regular". Table 4 presents the hit rates and distributions of errors in the *phone* simulations.

In general, these networks treat regulars which end in a dental as identities. However, many "dental final" regulars are mapped correctly and other regular stems are mapped as vowel changes or blended. Regular stems that conform to the characteristics of the vowel change class are often mapped as vowel changes, though again not all regular stems with vowel change characteristics are incorrectly mapped. There was a clear-cut advantage for the identity mappings in the *phone* condition. Given the well-defined sub-regularity that characterizes the identity class (all identity stems end in a dental), it is not surprising that the network is able to map this class successfully. The network is able to make use of the sub-regularities detectable in the input, however, it is also not indiscriminate in its categorization of verbs into classes on the basis of these sub-regularities (though they are of course a source of error). In several of the *phone* simulations, the identity class achieves optimal performance. However, *en route*, the mapping undergoes several reorganizations in which some identity stems are alternately treated as regular and vowel change stems *after* having been mapped correctly.

tab 4 (horisontal) full page

Finally, there was a moderate advantage for vowel change mappings in the *Phone* condition, particularly apparent in simulations 17, 19 and 27. In simulations 17 and 27, both the vowel change class and the identity class have relatively large token frequencies. In the *parent* condition, the lack of phonological sub-regularities permits the tendency towards identity mapping to "spill over" into the vowel change class. However, in the *phone* condition, the phonological regularity of the identity class restricts the application of identity mapping to items that possess these characteristics and hence reduces the level of interference with the vowel change class.

The provision of phonological constraints on class membership enables the network to construct a cleaner partitioning of the mapping problem space. Competition effects between classes diminish and overall performance improves. Nevertheless, *patterns* of learning are observed that are similar to those in the parent simulations. However, many errors do bear the stamp of the phonological structure of the input set. Competition between the verb classes, manifested as overgeneralizations, are observed where phonological information *is* or *is not* available to the system. The predominant error types for both sets of simulations are similar: Regulars, the most common error is identity mapping; Identities, the most common error is suffixation; Vowel changes, suffixation, identity mapping and blending, in that order. Similarly, blending errors in the identity class are absent in both the *parent* and *phone* simulations. In the *phone* simulations, temporary overgeneralizations are more predictable than in the *parent* simulations. Phonological sub-regularities are likely to be responsible for these emergent patterns of response.

Clearly, the *phone* simulations map input stems in light of the phonological information concerning class membership: overall hit rates improve, and the variety of errors is increasingly circumscribed. Many of the errors generated by these systems can be characterized in terms of the overapplication of a general strategy or "rule." However, these descriptions are generally insufficient to capture the diversity of behavior of the network. These networks increasingly come to resemble a rule-governed, categorical system as the constraints on the network (represented here though as *external* pattern constraints rather than *internal* architectural constraints) are tightened.

The constraining effect of the phonological sub-regularities in the *phone* simulations is particularly apparent in those simulations which otherwise give rise to substantial competition effects in the network

(compare *phones* 5, 17, 18 and 27 to the *parent* set). Phonological sub-regularities can, thus, serve to both *support and constrain* the type and token frequency effects observed throughout these simulations. Type/token frequency manipulations and phonological sub-regularities work together to support a high level of mapping performance across all classes. Just as the network manages to partition the arbitrary mappings in such a fashion that they appear immune to various parameter manipulations, the introduction of phonological sub-regularities in the other irregular classes results in a system which is increasingly impervious to type and token manipulations of the input vocabulary. However, type and token frequency effects do not disappear; rather, these effects are *modulated* by the internal structure of the sets of items that the network is required to process across learning.

6. Conclusions

This paper systematically explored the "acquisition" of mappings that are analogous to the English past tense by a simple 3 layer back-propagation network. The results revealed that class size and token frequency of the vocabulary used by the network crucially affected the degree to which characteristics of the different transformations "leaked" to verbs in other classes. The network was able to "overgeneralize" identity mapping and vowel change strategies, in addition to the suffixation procedure, given different input conditions. In some simulations, it was useful to describe the errors made by the network in terms of a general strategy or "rule," yet, the complexities inherent in the behavior of these systems did not typically warrant the use of such constructs. Overgeneralization errors, within any given simulation, were rarely restricted to a single type. In addition, different verb classes are susceptible to different types of errors. While the production of errors has been the focus here, it is important to note that several of these networks were able to master a substantial portion of this task. Given the appropriate input conditions, networks were able to "memorize" arbitrary forms at the same time that they were capturing the regularities in the other three classes. Yet, the mechanism guiding this memorization was the "same" as that guiding the rule-like overgeneralization behavior in the other classes. Likewise, phonological information resulted in complete mastery of the identity stems in several cases.

The degree to which these results are analogous to the acquisition patterns of children is not as yet totally clear. Some reports of children's past tense productions have provided examples of errors of vowel change or

identity "overgeneralizations" (e.g., Bybee/Slobin, 1982), yet the exact status of these as phenomena of acquisition which crucially require explanation has been subject to debate. In addition, some analyses suggest that children's production repertoires reflect a variety of general strategies which result in both correct and incorrect performance (Derwing/Baker, 1986; Marchman 1988). Children, like these networks, are not likely to be exclusively suffix generalizers, or identity mappers, but will produce several different types of errors in generating past tense forms throughout acquisition. Rule-based models explicate this phenomenon via the competition between two (or more) discrete and explicitly represented full-blown hypotheses which, at various points in development, undergo changes in how and when they are likely to apply (see Pinker/Prince, 1988). In these simulations, in contrast, probabilistic differences between individual mapping strategies are a natural by-product of learning (see Bates/Thal/Marchman, 1989) in that output fluctuations are the result of the implicit encoding of similarity relationships between the input stems in the weight matrix of the network. Here, high token frequencies tended to "localize" the zones of interference of mapping types while high type frequencies tend to extend them.

In addition, errors such as *ated* or *stooded* which blend two potential regularities in a single form were also observed (Kuczaj, 1977; Kuczaj, 1978). While these errors are also produced by children, they are much less frequent than the standard overgeneralization of "add -ed," and are likely to occur later in development. The absence of high-level organizing constructs such as "stem" and "suffix" may make PDP models particularly prone to this type of error, perhaps to a degree beyond the tolerance of an accurate model of children's acquisition (Pinker/Prince, 1988). However, while most simulations did indeed produce blends, they are relatively rare and predominate only in the vowel change class. Identity stems virtually never underwent blending. The introduction of phonological sub-regularities further restricted the occurrence of blending errors. Further analysis is required to outline the developmental priority of "pure" over blended overgeneralizations in these networks.

Within certain limits, these networks sometimes behaved as if they were doing what we know children do during the acquisition of morphological systems. Classic patterns of overgeneralization were elicited by manipulating token and type frequency, in networks that did and did not use phonological information to define class membership. However, the performance of these simulations must be considered to represent only an

approximation of the task required of a child who is learning language. Clearly, our representation of the input conditions is far from adequate, i.e., semantic information *must* play a role in the disambiguation of certain stem/past tense mappings. The degree to which these systems behave in ways that are reminiscent of the phenomena of acquisition are seen to reinforce the assumption that there is much to be gained from careful study of the nature and structure of input in the problem of language acquisition.

Literature

Bever T. G, ed. (1982): *Regressions in mental development: Basic phenomena and theories*. Lawrence Erlbaum Associates. Hillsdale, NJ.

Bever T. G., ed. (1989): The demons and the beast — Modular and nodular kinds of knowledge. In: *Interdisciplinary approaches to language: Essays in honor of S-Y. Kuroda*. C. Georgopoulos/R. Ishihara (Eds.). Kluwer Dordrecht. Holland.

Bowerman M. (1982): Reorganizational process in lexical and syntactic development. In: *Language Acquisition: The State of the Art*. E. Wanner/L. Gleitman (Eds.). Cambridge University Press. Cambridge, MA.

Bybee J. /Slobin D. I. (1982): Rules and schemas in the development and use of the English past tense. In: *Language 58*; 265–289.

Chomsky N. (1980): Rules and Representations. In: *Behavioral and Brain Sciences 3*; 1–61.

Churchland P. S./Sejnowski T. J. (1988): Perspectives on cognitive neuroscience. In: *Science 242*; 741–745.

Cybenko G. (1989): Approximations by superpositions of a sigmoidal function. In: *Mathematics of Control, Signals and Systems*.

Derwing B. L./Baker W. J. (1986): Assessing morphological development. In: *Language acquisition: Studies in first language development. Second edition*. P. Fletcher/M. Garman (Eds.). Cambridge University Press. Cambridge 1986

Elman J. L. (1988): Finding structure in time. *Center for Research in Language Technical report #8801*. University of California, San Diego.

Elman J. L./Zipser D. (1988): Learning the hidden structure of speech. In: *Journal of the Acoustical Society of America 83*; 1615–1626.

Fodor J./Pylyshyn Z. (1988): Connectionism and cognitive architecture: A critical analysis. In: *Cognition 28*; 3–71.

Hare H./Corina D./Cottrell G. W. A (1989): A connectionist perspective on prosodic structure. In: *Proceedings of the 15th Annual meeting of the Berkeley Linguistics Society.* Berkeley, CA.

Karmiloff-Smith A. (1979): *A functional approach to child language*. Cambridge University Press, Cambridge.

Karmiloff-Smith A. (1986): From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. In: *Cognition 23*; 95–147.

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

Kuczaj S. (1978): Children's judgements of grammatical and ungrammatical irregular past tense verbs. In: *Child Development 49*; 319–326.

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. In: *Cognition 28*; 195–247.

Lashley K. S. (1933): Integrative functions of the cerebral cortex. In: *Physiological Review 1933 13*; 1–42.

MacWhinney B. (1987): The competition model. In: *Mechanisms of language acquisition*. B.MacWhinney (Ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

MacWhinney B./Leinbach J./Taraban R./McDonald J. (1989): Language learning: Cues or rules? In: *Journal of Memory Language 28*; 255–277.

Maratsos M. (1988): Problems of Connectionism: Review of S. Pinker/J. Mehler (Eds.) Connections and Symbols. In: *Science 242*; 1316–1317.

Marchman V. (1984): Learning not to overgeneralize. *Papers and reports on child language development 24*, 69–74.

Marchman V. (1988): Rules and regularities in the acquisition of the English past tense. In: *Center for Research in Language Newsletter 2*.

Marslen-Wilson W. D./Welsh A. (1978): Processing interactions and lexical access during word recognition in continuous speech. In: *Cognitive Psychology 82*; 29–63.

Minsky M. L./Papert S. A. (1988): *Perceptrons: An Introduction to Computational Geometry. Expanded Edition*. MIT Press, Cambridge, MA.

Mozer A. (1988): A focused back-propagation algorithm for temporal pattern recognition. In: *Technical report #CRC-TR-88-3*. University of Toronto, Canada.

Palmer S. E. (1975): Visual perception and world knowledge. In: *LNR Research Group Explorations in cognition*. D. A. Norman/D. E. Rumelhart (Eds.). Freeman, San Fransisco.

Piaget J. (1953): *The origin of intelligence in the child*. Routledge & Kegan Paul, London.

Pinker S./Mehler J. (1988): *Connections and Symbols*. MIT Press, Cambridge, MA.

Pinker S./Prince A. (1988): On language and connectionism: Analysis of a parallel distributed processing model of language acquistion. In: *Cognition 28*; 73–193.

Plunkett K./Marchman V. (1989): Pattern association in a back propagation network: Implications for language acquisition. *Center for research in language, Technical Report # 8902*. University of California, San Diego.

Rosch E. (1973): On the internal structure of perceptual and semantic categories. In: *Cognitive development and the acquisition of language*. T. E. Moore (Ed.). Academic Press, New York, NY; 111–144.

Rumelhart D. E./Hinton G. E./Williams R. J. (1986): Learning internal representations by error propagation. In: *Parallel distributed processing: Explorations in the microstructure of cognition, 1: Foundations*. D. E. Rumelhart/J. L. McClelland (Eds.). PDP Research Group. MIT Press, Cambridge, MA; 318–362.

Rumelhart D. E./McClelland J. L. (1986): On learning the past tense of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition, 2: Psychological and Biological Models*. J. L. McClelland D. E./Rumelhart (Eds.). PDP Research Group. MIT Press, Cambridge, MA; 216–271.

Rumelhart D. E./McClelland J. L. (1987): Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. In: Mechanisms of Language Acquisition. B. MacWhinney (Ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

Seidenberg M. S./McClelland J. L. (1989): A distributed, developmental model of word recognition and naming. In: *Psychological Review 96*; 523–568.

Smolensky P. (1988a): The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Unpublished manuscript*.

Smolensky P. (1988): On the proper treatment of connectionism. In: *The Behavioral and Brain Sciences 11*; 1–23.

Strauss S. (1982): *U-Shaped Behavioral Growth*. Academic Press, New York.

Vygotsky L. (1962) *Thought and language*. MIT Press, Cambridge, MA.