(1993). In D. E. Meyer & S. Kornblum (Eds.) Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience. Cambridge MA: MIT Press. (pp. 655-688).

27 Toward a Theory of Information Processing in Graded, Random, and Interactive Networks

James L. McClelland

This chapter describes some initial steps toward a theory of the asymptotic and dynamic properties of systems in which information processing adheres to the principles of graded, gradual, random, interactive, and competitive processing. The goals for the theory are (1) to unify the results of a number of different experimental paradigms with a single theoretical framework; (2) to examine the conditions under which adherence to the principles will give rise to simple general regularities of information processing, and to examine whether the framework makes it possible to explain or predict cases in which these regularities will not hold; and (3) to examine interdependencies among the principles.

The principles, their motivation, and some of their general computational properties are described first. Then the three goals are discussed. Following this, two case studies are reviewed in which progress is made toward each of the three goals. A concluding discussion indicates directions for further development of the theory.

The principles are very general, but their examination will presuppose that information processing takes place in a parallel-distributed processing (PDP) system (Rumelhart, Hinton, and McClelland 1986). A *PDP system* is simply a system in which processing occurs through the interactions of a large number of simple, interconnected processing elements called units. These elements may be organized into modules, each containing a number of units; sets of modules may be organized into pathways, each containing a set of interconnected modules. Pathways may overlap, in that they may contain modules in common. Processing in a PDP system occurs by the propagation of activation among the units, via weighted connections. The knowledge that governs processing is stored in the weights of the connections, and the effects of experience on information processing are captured by changes to the connection weights.

The PDP framework is an extremely broad framework and can be used to address a very wide range of different modeling goals, from efforts to capture the detailed properties of specific neural circuits to efforts to solve problems in artificial intelligence that have not yielded to more traditional symbolic approaches. The PDP framework has also been used for psychological modeling, and it has been very useful in this regard; but it is really best construed as

a framework providing tools with which to construct a theory, and not as a theory in and of itself. For the general framework is so broad that it does not provide much guidance or constraint without further assumptions.

27.1 PRINCIPLES

The following list of principles begins to provide such a constraining framework. While each of the principles has been used in previous work by the present author, none of the principles can easily be attributed to any particular source, since each is in extremely common use.

1. The activation of each unit is a graded, sigmoid function of its summed input.
2. Activation propagates gradually in time.
3. Between-module connections are mutual and excitatory, so that processing is interactive.
4. Within-module connections are mutual and inhibitory, so that processing is competitive.
5. The activation process is intrinsically variable.

It should be stressed that this set of principles is provisional. They should be viewed as a starting place and guide for research. No doubt there are other principles in addition to these, and no doubt, some or all of the principles will require further refinement.

Taken together with the basic characteristics of PDP, these principles define a modeling framework called the GRAIN model. GRAIN stands for graded, random, adaptive, interactive, (nonlinear) network. The framework encompasses adaptation or learning as well as the principles enumerated above, but the principles of adaptation are not yet fully clear and will not be considered further here. For a discussion of learning in GRAIN networks, see Movellan and McClelland (1991) and Cohen, Servan-Schreiber, and McClelland (in press).

Also missing from the list above are principles of representation. Although distributed representations have considerable advantages (see McClelland and Rumelhart 1985; Hinton, McClelland, and Rumelhart 1986) the particular models that will be the focus of interest in what follows, and the motivations for them, are couched in localist terms. That is, each processing unit stands for a specific choice that might be made in interpreting an input at a particular level of description. This feature arises not as a matter of principle but as a simplification. Of course, further work is required to examine the extent to which the results obtained for localist networks actually transfer to systems using distributed representations.

In this connection it is worth noting that models based on the above principles should not be construed as models of the neuronal processes underlying cognition. The units do not correspond to individual neurons, nor do the connections correspond to individual synapses. Rather, activations of units represent representational states of a processing system and connections capture constraints that hold among these representational states. A discussion of the relation between such models cast at this rather cognitive level and models of the underlying neural substrate may be found in Smolensky (1986).

The reader will note that the principles are stated in qualitative terms, without specific detailed quantitative assumptions. While particular simulation models must be formulated in terms of specific quantitative assumptions, it can often be shown that these details are relatively unimportant. It does not appear to matter, for example, what the exact form of the graded sigmoid function is, or whether the intrinsic noise is Gaussian or uniformly distributed in a bounded interval.

Discussion of the Principles

The motivations for the principles are complex and interdependent, and to do full justice to each would require much more space than is available here. In what follows, direct and basic psychological motivations are given for each principle. Sometimes, however, the motivation for one principle arises primarily from its interaction with others. These interactions are addressed more fully in the later parts of this chapter, but forward pointers to some of the most important interactions are given at the end of this section. In general, the principles and their motivations are attributable to a number of sources; no claim for priority is intended by listing them here.

In what follows some basic theoretical results concerning networks that adhere to the principles are also noted. More details can be found in a number of sources; the most up-to-date review is given in Hertz, Krogh, and Palmer (1990).

Graded, Sigmoid Activation Function The use of a graded activation function of summed input allows a variable to exert a continuous or graded influence on processing outcomes, while leaving these outcomes open to other influences. Graded influences on cognitive outcomes are ubiquitous. To mention just one key example, category membership appears to be a graded function of similarity to typical exemplars. Openness to additional influences allows for factors such as contextual or attentional inputs to influence the outcome of processing. Graded influences allow for collaboration and competition of cues, and play a role in a number of basic approaches to processing (Oden and Massaro 1978; MacWhinney 1987).

A sigmoid activation function is simply a continuous activation function that is monotonically increasing, has a single point of inflection, and levels off at both extremes (fig. 27.1). Such a function is used to characterize the asymptotic activation produced by a fixed net input to a unit. The motivation for the use of a sigmoid activation function, as opposed to a completely linear activation function, is basic: multilayer networks that use linear activation functions are computationally trivial. They cannot compute anything that

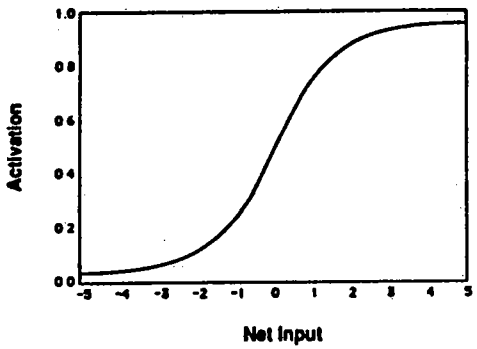

Figure 27.1 A sigmoid function of the type used in many connectionist models.

cannot be computed with a single layer of units, and what can be computed by a single layer of units is very limited indeed (Minsky and Pappert 1969).

Furthermore, linear networks with bidirectional connections, as required by the interactivity principle, can easily exhibit explosive growth of activations. The use of a sigmoid activation function prevents this. On the other hand, the sigmoid is a very simple function; it is, for example, monotonic in each of its inputs, independently of all the others (Williams 1986).

Once a sigmoid function is used, the effect of an input on the outcome of processing depends on other inputs. Each input (excitatory or inhibitory) tends to push the activation of each unit in a particular direction, but the magnitude of the effect depends critically on the other inputs to the unit. This allows us to understand, for example, how a cue can influence response probabilities when other cues are relatively neutral but fail to have much influence when other cues strongly favor one particular response.

Once sigmoid functions are used in multilayer systems, arbitrarily complex patterns of activation through time can arise. Further constraints are crucial. The use of bidirectional connections helps a great deal, as will be discussed below.

Gradual Propagation of Activation The assumption of gradual propagation arose in the context of efforts to study the dynamics of information processing and of contextual influences on processing (McClelland 1979; Rumelhart and McClelland 1981). One key observation here is that it often takes subjects less time to recognize a stimulus when they use more information to do so. Thus recognition may be both more accurate and faster when a stimulus occurs in a predictive context (e.g., a word appearing in a sentence context) than when it appears alone. In many such experiments (e.g., McClelland and O'Regan 1981) context alone is not sufficient for recognition, so information from both the stimulus and the context must be used. A very natural and direct way to account for such effects is to assume that information about both the stimulus and the context accumulates gradually over time, so that when both sources of information are accumulating at the same time a response threshold is reached more quickly.

Interactive Processing The idea that processing is interactive arose in the author's thinking from Rumelhart's (1977) seminal paper, "The Interactive Model of Reading." (Grossberg 1978a also explored this idea, though mostly for different reasons, at around the same time.) The key point of Rumelhart (1977) was to argue that in reading (as in speech perception and other perceptual processes) decisions at each level—in this case feature, letter, word identity, word meaning, syntactic structure, role assignment, anaphoric reference, etc.—reflected influences from all other levels. McClelland and Rumelhart (1981; Rumelhart and McClelland 1981, 1982) developed a model that applied this assumption to visual letter recognition. This work was motivated by the fact that the perception of a letter in a word is facilitated more when it occurs in context (Reicher 1969) than when it occurs in isolation. The account for this was simply that partial activations of context and target letters could give rise to activations of words and these in turn could feed back activation to the letters, increasing their activation compared to cases in which a single letter was presented in isolation. Elman and McClelland (1988) confirmed a key prediction of this same model in a speech perception context by showing that phonemes whose identification depended on lexical context could trigger coarticulatory influences on the identification of other phonemes. Such influences have long been studied in the speech perception literature, and they are generally interpreted as operating at or below the phoneme level. The finding that these coarticulatory influences could be triggered by lexical context was predicted from the principle of interactivity, and contradicts feed forward models in which lexical influences do not feed back to the phoneme level.

Mutual Competition Competitive interactions among units have been used by many investigators (e.g., Feldman and Ballard 1982; Grossberg 1976, 1978a,b). They allow contrast enhancement and the suppression of weak activations. They also allow the activation of each alternative to be influenced by the extent to which there is input that favors other alternatives. Response choices reflect this weighing of alternatives relative to others (Luce 1959), and mutual inhibition can implement this relative weighting process (Grossberg 1978b). In McClelland and Rumelhart (1981), the use of this idea at each of several levels of processing was motivated originally by the desire to distribute decision making throughout information processing systems, rather than to centralize this function in a limited capacity executive.

The general principles of interactivity and competition are consistent with a specific network constraint, called symmetry, in which for all pairs of units A and B, the connection from A to B has the same value as the connection from B to A. Obviously interactive and competitive networks need not be strictly symmetric, but they can be. Symmetric networks have a very important property, which is that there exists some function (known as a Liapunov

function) whose value changes monotonically as processing proceeds through time. Hopfield (1982) was the first to point this out. His symmetric networks monotonically decrease the value of what Hopfield called an energy function; alternatively the minus sign may be removed, and the function, which is now monotonically increasing, may be called a goodness function (Rumelhart et al. 1986). Choosing the latter formulation, it can be shown that symmetric networks always proceed uphill in goodness as processing continues, where goodness is defined as the extent to which the activations in the network satisfy the constraints represented by the inputs to the network and the connection weights (see Rumelhart et al. 1986 for a full discussion). When activations are bounded, as they are in GRAIN networks, the activations and thus the goodness eventually stop increasing as processing continues. This point will have higher goodness than any neighboring points. Such a point is called a local maximum.

The fact that networks are monotonically adjusting the value of some function as the activation process proceeds from a starting point does not mean that the activation of each unit will vary monotonically with time. Such nonmonotonicities in the time course of activation arise from the changing pattern of activation surrounding each unit, sometimes causing units that receive net excitation at the beginning of processing to receive net inhibition at a later point. Examples of this were discussed in McClelland and Rumelhart (1981) and a different example will be considered again below.

Variability That human performance is variable cannot really be subject to much doubt, and in fact most models of information processing dynamics are stochastic models (see Luce 1986 for a review). Some authors (e.g., Anderson 1991) have stressed that manifest variability in reaction times across different test items (e.g., in a recall experiment) may in fact reflect differences among items rather than actual variability in the processing itself, and there are surely many other factors (mood, context, motor preparation, accommodation of the eye, etc.) that introduce trial-to-trial variation. What is of interest here is that incorporating some source of variability, either in the input to processing or intrinsic to processing activity itself, improves our ability to model human information processing, relative to a deterministic model in which variability is introduced only at the response selection stage. This matter will be considered at great length in a subsequent section. Here we consider a very general characteristic of stochastic symmetric networks.

Stochastic symmetric networks, like deterministic ones, are minimizing a Liapunov function, but this function characterizes distributions of states rather than individual states. Over time they settle to a particular distribution over possible states of activation, independent of the starting point of the settling process (this is a basic general fact; see Movellan and McClelland 1991 for a fuller discussion). Note that the actual pattern of activation does not stabilize; what stabilizes is the probability that the network will be in each possible activation state. At the beginning of settling, the distribution of states of the network will be determined largely by the initial pattern of activation. But as time goes on, the distribution eventually becomes independent of the starting point. Note that initial patterns of activation should not be confused with fixed inputs that remain on throughout the settling process. These do affect the equilibrium distribution, making some patterns more probable at the expense of others.

It is actually possible to write down the equilibrium probability distributions of some stochastic symmetric networks as a function of their fixed inputs and of the connection strengths. This fact will be used below in discussing derivations of asymptotic choice probabilities.

27.2 GOALS FOR A THEORY OF INFORMATION PROCESSING

There are many possible goals that we may have for a theory of information processing. For the present we will focus on the three goals that were mentioned at the beginning of this chapter: (1) synthesis of the results of a range of different experimental paradigms, (2) exploration of the conditions under which systems that adhere to the principles can be characterized by simple general laws, and (3) analysis of the interdependencies among the principles.

Synthesis

One goal for a theory of information processing is to bring together in a single model results from a variety of different paradigms. The three main classes of paradigms of interest here will be designated (1) the asymptotic choice paradigm, (2) the reaction-time paradigm, and (3) the time-accuracy paradigm. Though the paradigms will be familiar to most readers, a brief review of their characteristics and their relation to the present approach will be worthwhile.

The asymptotic choice paradigm encompasses experiments in which subjects are asked to make choice responses (or yes-no responses) without time pressure, and the assumption is that they wait until there is nothing further to be gained from waiting. The dependent measure is the distribution of responses over alternatives. Such situations are the easiest to model, since as already noted it is often possible in connectionist systems to characterize the distributions of asymptotic states mathematically as functions of inputs and connections.

The reaction-time paradigm encompasses experiments in which subjects are asked to make responses as rapidly as possible without sacrificing accuracy. Ordinarily the main dependent variable is simply mean reaction time. From the point of view of graded, stochastic systems that adhere to the principles outlined above, this paradigm is the least satisfactory. First of all, while models based on the above principles can in principle model mean reaction times, there is one main difficulty. This is simply that in systems where there is a gradual accumulation of graded information, the instruction to respond as rapidly as possible without sacrificing accuracy is not well defined (Wickelgren 1977;

Pachella 1974; Ratcliff 1978). All else being equal, it is very commonly observed that accuracy is lowest in conditions producing the longest reaction times. While this fact arises naturally from some continuous models (see Ratcliff 1978) in which a fixed evidence criterion is used, it is also very likely that the criteria themselves may slip, as the subject becomes impatient to respond. To model the data requires complex assumptions about criteria which can be both extraneous to the theory and difficult to check.

It is true that some researchers (e.g., Ratcliff 1978) make extensive use of dependent measures other than mean reaction time. These include error rates and properties of reaction-time distributions that characterize their form and spread. When such properties are available for both correct and error responses, for each of several conditions, the overall pattern of data certainly imposes a considerable degree of constraint. However even here assumptions about criteria remain, and considerable information about the actual time course of processing is lost.

The time-accuracy paradigm encompasses a variety of procedures that attempt to relate accuracy to elapsed time since the onset of a stimulus. This can be done in a variety of ways. Sometimes different deadlines are used in different blocks of trials (Pachella 1974). Sometimes a response signal is used, and subjects are trained to respond very shortly after the occurrence of the signal (Wickelgren 1977). In these cases, accuracy is typically plotted as a function of mean reaction time in each deadline or response signal condition. Such curves are called time-accuracy curves. In other variants, subjects are simply told to go so fast that overall accuracy falls to something like 70–80% (Lappin and Disch 1973; Gratton et al. 1988). This often induces a broad distribution of reaction times for each experimental condition, with relatively low accuracy for short reaction times. The RT distribution can be divided into bins, and accuracy for responses falling in each bin can be plotted, yielding what Gratton et al have called a "conditional accuracy function." Although time-accuracy curves and conditional accuracy functions are not the same, both can provide information about the accumulation of information through the changing distribution of responses over time. Thus both are useful tools for studying graded, stochastic processes.

A theory of information processing dynamics like the present one ought to be able to provide insight not only into asymptotic choice behavior, but also into the time course of change in response-choice probabilities. Although the time course of nonlinear stochastic dynamical systems is known to be difficult to analyze mathematically (Cox and Miller 1965), the hope is that simulations, supplemented where possible by mathematical analyses of boundary conditions, can provide some insight into various qualitative aspects of these systems.

Because of the ubiquity of reaction-time experiments, any reasonable theory must address results from this paradigm as well. However in the present chapter there is room to cover only a limited range of results, so the focus will be on recent analytic progress characterizing effects of variables on asymptotic choice probabilities and on simulations addressing aspects of the time course of processing as revealed through time-accuracy studies.

Complex Systems, Simple Laws

The principles stated above are remarkably simple, but processing systems that adhere to them can be very complex, both in terms of their structure and in terms of their behavior. Yet often the data produced in information processing tasks can be captured by very simple laws. This chapter will be concerned with two examples.

Morton's Independence Law for Effects of Context and Stimulus Information in Perception It has been repeatedly observed that context and stimulus information produce additive, or independent, effects on the z-transform of stimulus identification response probabilities. The generality of this pattern was first noted by Morton (1969), and has been stressed by Massaro (e.g., Massaro 1989). It is illustrated in graphical form in figure 27.2.

Wickelgren's Law for Time-Accuracy Curves In studies where response accuracy is measured at different times after stimulus onset, a simple general pattern is typically observed (fig. 27.3). Responses that occur immediately after stimulus onset are essentially random, and unrelated to the stimulus. At a later point, accuracy begins to rise rapidly above chance, then gradually levels off. The overall pattern of performance can generally be described as a shifted exponential approach to asymptote, as Wickelgren (1977) was the first to point out. These regularities will be called Morton's independence law and

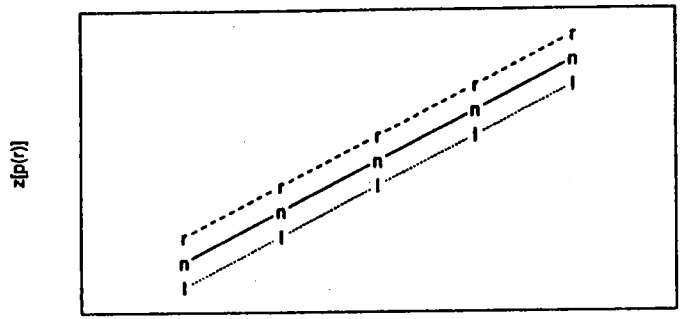

Figure 27.2 Idealized pattern of results expected in the Massaro and Cohen experiment if the independence law holds. The stimulus conditions are ordered from *l*-like to *r*-like on the x-axis. The y-axis represents the z-score of the probability of the *r* response. This is plotted for each of three context conditions. Curves are labeled *r*, *l*, and *n* for contexts favoring *r*, *l*, and *n*, respectively. Reprinted with permission from McClelland (1991).

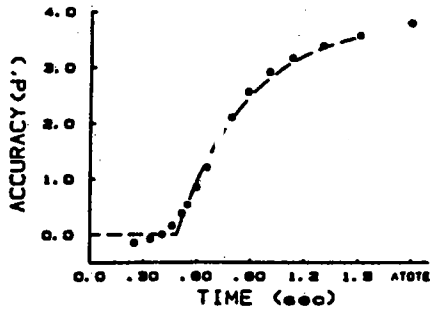

Figure 27.3 The dashed line depicts the shifted-exponential curve proposed by Wickelgren (1977) that characterizes the time-accuracy curves that have been obtained in a large number of different experiments. A composite of data from a number of different experiments demonstrates the close correspondence. Reprinted with permission from McClelland (1979).

Wickelgren's law for ease of reference. In calling them laws, no claim is made that either actually always holds, and indeed we shall be concerned with cases where they do not hold; they are, however, robust empirical regularities that do hold under a wide range of circumstances.

Typically, such simple regularities in information processing have led to simple models. Morton's (1969) model of context effects in perception, and the more recent model of Oden and Massaro (1978) are clearly cases in point. In the Oden and Massaro model, for example, it is assumed that independent sources of evidence are evaluated independently, then combined according to a simple combination rule, to yield response strengths for each alternative. These response strengths then enter into a decision process which selects an alternative probabilistically based on the Luce (1959) choice rule.

One of the central goals of the present work is to understand whether, and under what circumstances, these (and eventually other) simple laws might arise from processing systems that adhere to the principles described above. A related and equally important goal is to delineate circumstances in which these simple patterns of data break down. The hope is that it will be possible to see just what additional constraints beyond the basic principles themselves are needed for the simple laws to hold, and what the consequences of violations of these assumptions may be.

The goal of subsuming empirical regularities under a set of general principles is one that the present work shares with Newell's (1990) work on unified theories of cognition. The goal of going beyond merely subsuming these regularities to giving some account of cases where they hold and do not hold is curiously absent from Newell's approach. The view taken here is that the utility of the general principles is perhaps most strongly established when they can lead us beyond general laws that usually hold, to an understanding of how they might fail and what it might mean when this happens. We will see that two sources of information do not always exert independent effects on performance, and that time-accuracy curves are not even always monotonic; and we will see that these cases can be understood in terms of violations of architectural constraints on the structure of GRAIN networks.

Interdependencies of the Principles

As shall become evident below, there are cases in which models that adhere to only four of the five principles produce poor fits to data, but models that adhere to all five produce excellent fits. This kind of discovery argues strongly in favor of the effort to articulate and evaluate an entire *set* of principles, rather than to try to evaluate the validity of the principles one-by-one (Newell 1973, 1990). For if there are interdependencies, any finding we might obtain about the validity of one of the principles will change with changes in other assumptions. For this reason, an effort to understand interdependencies among the principles is among the central goals of the theory.

27.3 CASE STUDIES

In this section, two case studies are reviewed. In each case we will see that GRAIN networks exhibit correspondence to well-known general laws—but only under certain boundary conditions. In the first case study, the analysis is somewhat more developed. Here it has been possible to use mathematics to analyze asymptotic choice performance, to relate the parameters of the GRAIN model to parameters of classical models of asymptotic choice performance, and to generate and confirm a prediction that arises from the mathematical analysis for a case in which the general laws will not hold. The second case study illustrates one future direction for the work with simulation studies of performance in the time-accuracy paradigm. Here we find that the GRAIN model can produce curves that approximate Wickelgren's time-accuracy law, while also providing a framework for developing a plausible account of one case in which this law fails.

27.3.1 ASYMPTOTIC PERFORMANCE IN PERCEPTUAL IDENTIFICATION TASKS

The Problem

A basic assumption of the GRAIN model is the principle of interactivity, or bidirectional influence between mutually consistent units at different processing levels. However, Massaro (1989) called this principle into question. He pointed out that the interactive activation model (McClelland and Rumelhart 1981), which embodied this principle, violated the independence law mentioned above. Here we consider the independence law in the context of a particular experiment: the determination of the identity of an ambiguous consonant under the influence of perceptual and contextual information. In this experiment (Massaro 1989) perceptual information was varied by creating seven tokens of an ambiguous segment on a continuum from /l/ to /r/.

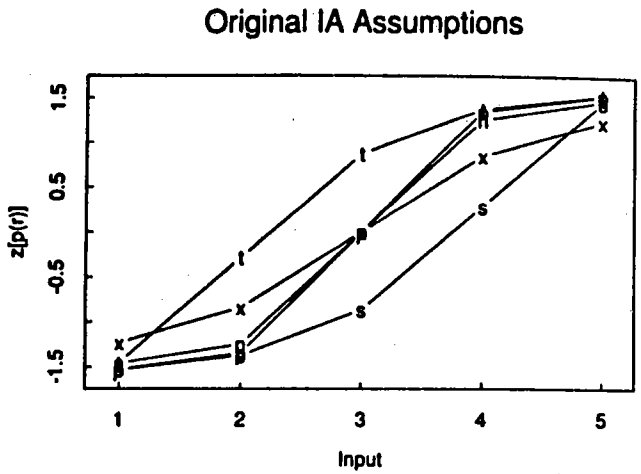

Figure 27.4 Simulation results from the interactive activation network shown in figure 27.5. Reprinted with permission from McClelland (1991).

Context was varied by presenting each token in a context in which both /r/ and /l/ can appear (/p_i/ as in *pre* or *plea*), a context where /r/ can appear (/t_i/ as in *tree*), a context where /l/ can appear (/s_i/ as in many words such as *sleep, sleet,* etc.) and finally a context in which neither actually occurs in English (/v_i/). Subjects must simply identify the ambiguous token as /r/ or /l/, without time pressure.

The independence law states that we should be able to space the seven input conditions along the x-axis of a graph (representing a log-likelihood continuum from /l/ to /r/) in such a way that the results from the 28 conditions of the experiment form four (straight and) parallel lines, one for all seven of the points in each of the four contexts (fig. 27.2). In fact the data from a large number of similar experiments can indeed be fit fairly closely in just this way.

What Massaro noted was that the interactive activation model as originally formulated failed to produce the required straight lines. The actual curves produced in a simulation of a simplified network appropriate for modeling the Massaro data is shown in figure 27.4. The network is shown in figure 27.5. It consists of two target units, one for /r/ and one for /l/; three context-input units, for the initial phonemes /s/, /p/, and /t/ from the Thompson and Massaro experiment; and four higher-level units, one for each of the legal combinations of context and target letters. Thus there are higher-level units for *pr, pl, tr,* and *sl* but not for *sr, tl, vr* or *vl*. (A fourth context input unit for /v/ would not have been helpful since /v/ is not connected to anything.) Five steps along the /r/–/l/ continuum were considered, varying the input to the /r/ unit from 0.3 to 0.7 in steps of 0.1. The input to the /l/ unit was set equal to one minus the input to the /r/ unit. The four context conditions were simulated by supplying an input of 1.0 to the appropriate context unit (or to

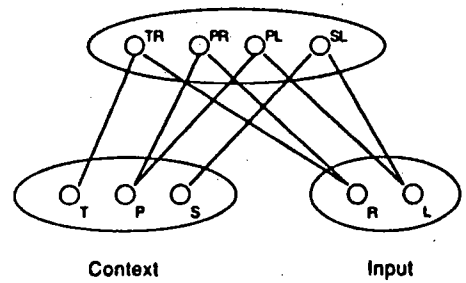

Figure 27.5 The network used in the simulations of the joint effects of context and stimulus information in phoneme identification. Reprinted with permission from McClelland (1991). Lines indicate bidirectional excitatory connections. All units within the same enclosed region are mutually inhibitory.

no unit at all, in the case of the /v/). Each simulated trial began with activations of all units set to rest (−.1), and appropriate inputs were turned on and left on.

Processing occurred according to the specific activation assumptions of the interactive activation model (using the iac program of McClelland and Rumelhart 1988). These assumptions are consistent with the GRAIN model in most respects. The activation function is a graded sigmoid function; gradual propagation of activation is assumed; between-level connections are bidirectional and excitatory, and within level connections are bidirectional and inhibitory. But there is no variability; processing itself is completely deterministic. Probabilistic performance is "tacked on at the end," to use a phrase from Luce (1986): At the end of 60 cycles activations of the /r/ and /l/ units were used to determine response probabilities according to the Luce choice rule applied to the exponential of the activation of each unit, as in McClelland and Rumelhart (1981).

As figure 27.4 indicates, the interactive activation model did not produce parallel lines in this simulation. While it is possible to space the points on the x axis to make *one* of the curves a straight line, no transformation can make all the lines parallel, since some of them actually cross.

Massaro's (1989) claim was not simply that the interactive activation model failed in particular cases to produce the right pattern of results. Rather, the claim was that the principle of interactivity itself was incompatible with independence. He based this claim on the following intuitively appealing argument. The independence law states that each source of information makes an independent (additive) contribution to the evidence for or against each alternative. But in the interactive activation model, the effects of context and direct stimulus information on activations of letter detector units are not independent. That is, the effects of one source of evidence—direct stimulus input—are influenced by the presence of bidirectional connections between the letter and word level units. Bottom-up input can activate a letter unit, which can in turn feed upward and activate a word unit, and the word unit will in turn send activation back to the letter unit. This process seems to be sending

back from the word level some of the activation that came up to it from the letter level. Further, and this is where independence is violated, the extent of the activation that is fed back depends on context; the context influences the activations of word level units, and if a particular context activates supporters of one target alternative rather than another, only the former will have the benefit of recurrent activation; the latter will not get this benefit. Thus, the context appears to modulate the extent of activation a letter gets from stimulus input, thereby violating the independence assumption.

The Role of Variability

Interestingly, this intuitively appealing argument only applies to the deterministic version of the interactive activation model studied by McClelland and Rumelhart; it does not hold when there is variability, either in the input or in the processing itself (McClelland 1991). The presence of variability actually permits a simplification of the model in one respect: When there is variability, responses can be made simply by settling until equilibrium is reached and then picking the most active alternative. Under these conditions, the network itself is in a real sense choosing the response, and the probabilistic character of these choices arises from the variability. Under these conditions, interactive activation models do behave in accordance with Morton's independence law.

Simulation results establishing the independence pattern for the network shown in figure 27.5 are shown in figures 27.6a and 27.6b. For figure 27.6a, variability was added to the input to the network. Specifically, the input to the /r/ unit was perturbed by a Gaussian random noise with 0 mean and standard deviation 0.14. Again the input to the /l/ unit was one minus the (perturbed) input to the /r/ unit. Each trial was run just as before; processing was completely deterministic. All of the variability was in the input, which stayed constant within a given trial. Response choices were made by simply selecting the alternative with the largest activation at the end of 60 cycles (other values were used in other runs with indistinguishable results). Of course, the variability in the input made performance probabilistic, and many runs are needed to establish the actual probabilities. The graphs are based on 10,000 independent simulation trials per data point.

For figure 27.6b, the external input to each unit was fixed as in the initial simulation; this time all of the variability was intrinsic to the processing in the network. On each time step, a sample of noise from a Gaussian distribution with mean 0 and standard deviation 0.14 was added into the net input of each unit. Again, after 60 cycles of processing, the network's response was determined by selecting the most active response unit.

Analysis The fact that the model conformed to the independence law can be established mathematically for the case where the variability occurred in the input. A formal presentation can be found in McClelland (1991). For the present, the following informal analysis is given. In the case under consideration, the network itself remains completely deterministic. This means that

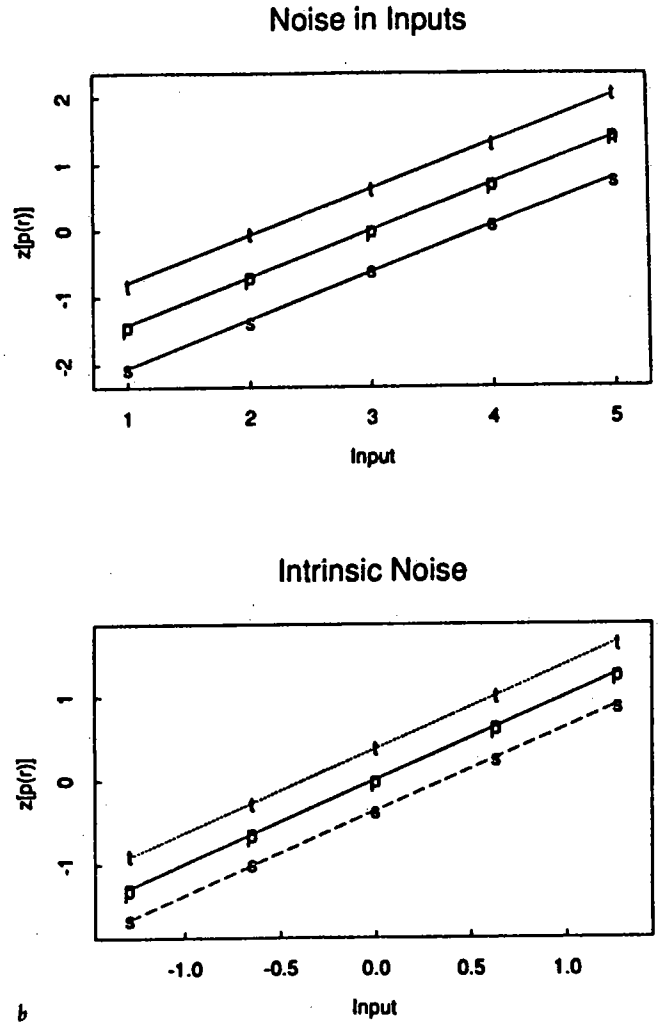

Figure 27.6 Simulations based on the network shown in figure 27.5 with perturbed inputs (*a*) and intrinsic variability (*b*). Reprinted with permission from McClelland (1991).

each particular choice of inputs (context plus direct input to /r/ and /l/) always produces the same output. Consider first a series of trials, without any contextual input, but with input to /l/ (denoted $i(l)$) increasing across trials in small steps from 0 to 1.0, while input to /r/ is decreasing correspondingly so that $i(l) + i(r) = 1$. The network is symmetric with respect to /l/ and /r/ in this case and so when $i(l) < i(r)$, $a(l)$ will be less than $a(r)$. So in this case, the network will choose /l/ whenever $i(l) > i(r)$, and will choose /r/ otherwise.

Now consider what would happen with some contextual input $c(a)$, and let us suppose that the contextual input in question favors /r/. Then in this case, when $i(r) = i(l)$, and even when $i(r)$ is slightly less than $i(l)$, $a(r)$ will be greater than $a(l)$, and so the subject will choose /r/. However, as we increase $i(l)$ and correspondingly decrease $i(r)$ there will come a point at which the disparity in the input is enough to overcome the differential effect of the context. At this point, $a(l)$ will be greater than $a(r)$ after settling, and the response chosen would be /l/. A similar argument can be given for another context, $c(b)$, that favors alternative /l/.

In short, when the network itself is deterministic, each context establishes a different cut-point along an axis that represents $i(l)-i(r)$. All values of $i(l)-i(r)$ that are less than this cut-point result in the /l/ response; all those above the cut-point result in /r/. Each different context establishes a different cut-point.

Figure 27.7 shows the distributions of $i(l)-i(r)$ for three different stimulus input conditions with Gaussian noise in the input. The figure also shows three cut-points, one for a context favoring /r/, one for a context favoring /l/, and one for a neutral context. As the figure illustrates, the effect of context is simply to shift the criterion along the $i(l)-i(r)$ continuum. Now, the z-score of the probability of choosing the /l/ response is the distance, in units of the standard deviation of the noise, from the mean value of $i(l)-i(r)$ to the cut-point. Obviously changes in the input ($i(l)-i(r)$) or in the criterion, have independent, additive effects on this distance. Thus the model directly captures Morton's law.

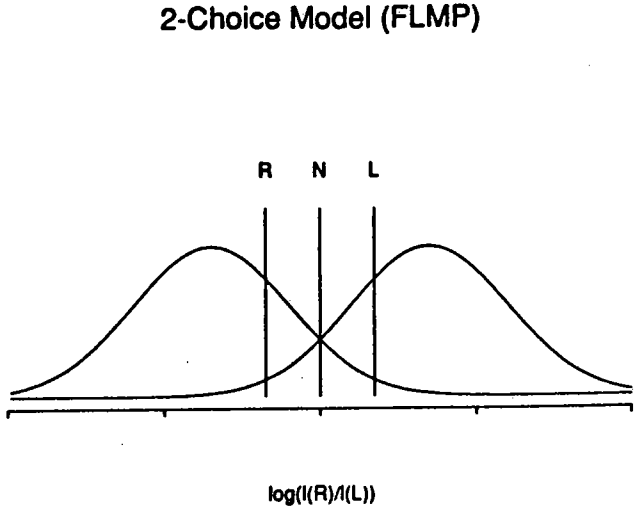

Figure 27.7 Distributions of $i(l)-i(r)$ for two different stimulus input conditions with Gaussian noise in the input, together with cut-points imposed by the interactive activation process for three different contexts. The cut-point labeled *R* represents a context favoring *R*; *L* represents a context favoring *L*; and *N* represents a neutral context. Reprinted with permission from McClelland (1991).

This informal argument showing that the network of figure 27.5 captures Morton's law when variability is introduced in the inputs to an interactive activation network is presented more formally in McClelland (1991). However, for the case where the noise is intrinsic to processing, the result was established by simulation only. A formal derivation of the result was possible if the particular processing assumptions of the interactive activation model were replaced by those of the Boltzmann machine (Hinton and Sejnowski 1986).

The Boltzmann machine is inherently a stochastic and interactive model, but it makes use of binary, rather than graded, activations. Units are updated one at a time in random order, and activations are set to 1 with a probability given by the logistic function of the net input to each unit:

$$p(a_i = 1) = 1/(1 + e^{-net_i/T}).$$

This is the actual sigmoid function shown in figure 27.1: now it is being used to give the probability that the activation is 1, rather than the actual activation of the unit.

For present purposes T, a parameter called temperature, is taken to be a fixed parameter that simply scales the magnitude of the net input. As before, trials are run by presenting input plus context, leaving them on until equilibrium is reached; and then selecting, at some particular time t, the most active member of the alternative set as the network's response. There is one minor complication: In Boltzmann machines, units representing particular alternatives are either on or off; therefore when a choice must be made, a tie is possible. In case of a tie, it is necessary to resample at a later time, repeating until no tie occurs.

The Independence Constraint on Architecture

When the above choice procedure is used, it can be shown that Boltzmann machines (symmetric, stochastic networks using the above binary update function) will conform to Morton's law as long as the network adheres to a particular architectural constraint. This constraint can be stated as follows:

1. The network must be partitionable into three sets of units: one set representing the alternatives, one set representing the stimulus input to the alternatives, and one set representing the context relevant to the alternatives.

2. No connections are allowed between any of the units in the input set and any of the units in the context set. The only interactions among the sets occurs by way of the units representing the alternatives.

each particular choice of inputs (context plus direct input to /r/ and /l/) always produces the same output. Consider first a series of trials, without any contextual input, but with input to /l/ (denoted $i(l)$) increasing across trials in small steps from 0 to 1.0, while input to /r/ is decreasing correspondingly so that $i(l) + i(r) = 1$. The network is symmetric with respect to /l/ and /r/ in this case and so when $i(l) < i(r)$, $a(l)$ will be less than $a(r)$. So in this case, the network will choose /l/ whenever $i(l) > i(r)$, and will choose /r/ otherwise.

Now consider what would happen with some contextual input $c(a)$, and let us suppose that the contextual input in question favors /r/. Then in this case, when $i(r) = i(l)$, and even when $i(r)$ is slightly less than $i(l)$, $a(r)$ will be greater than $a(l)$, and so the subject will choose /r/. However, as we increase $i(l)$ and correspondingly decrease $i(r)$ there will come a point at which the disparity in the input is enough to overcome the differential effect of the context. At this point, $a(l)$ will be greater than $a(r)$ after settling, and the response chosen would be /l/. A similar argument can be given for another context, $c(b)$, that favors alternative /l/.

In short, when the network itself is deterministic, each context establishes a different cut-point along an axis that represents $i(l)-i(r)$. All values of $i(l)-i(r)$ that are less than this cut-point result in the /l/ response; all those above the cut-point result in /r/. Each different context establishes a different cut-point.

Figure 27.7 shows the distributions of $i(l)-i(r)$ for three different stimulus input conditions with Gaussian noise in the input. The figure also shows three cut-points, one for a context favoring /r/, one for a context favoring /l/, and one for a neutral context. As the figure illustrates, the effect of context is simply to shift the criterion along the $i(l)-i(r)$ continuum. Now, the z-score of the probability of choosing the /l/ response is the distance, in units of the standard deviation of the noise, from the mean value of $i(l)-i(r)$ to the cut-point. Obviously changes in the input ($i(l)-i(r)$) or in the criterion, have independent, additive effects on this distance. Thus the model directly captures Morton's law.

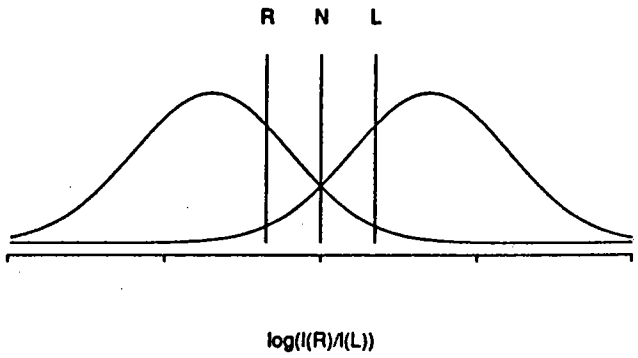

Figure 27.7 Distributions of $i(l)-i(r)$ for two different stimulus input conditions with Gaussian noise in the input, together with cut-points imposed by the interactive activation process for three different contexts. The cut-point labeled *R* represents a context favoring *R*; *L* represents a context favoring *L*; and *N* represents a neutral context. Reprinted with permission from McClelland (1991).

This informal argument showing that the network of figure 27.5 captures Morton's law when variability is introduced in the inputs to an interactive activation network is presented more formally in McClelland (1991). However, for the case where the noise is intrinsic to processing, the result was established by simulation only. A formal derivation of the result was possible if the particular processing assumptions of the interactive activation model were replaced by those of the Boltzmann machine (Hinton and Sejnowski 1986).

The Boltzmann machine is inherently a stochastic and interactive model, but it makes use of binary, rather than graded, activations. Units are updated one at a time in random order, and activations are set to 1 with a probability given by the logistic function of the net input to each unit:

$$p(a_i = 1) = 1/(1 + e^{-net_i/T}).$$

This is the actual sigmoid function shown in figure 27.1; now it is being used to give the probability that the activation is 1, rather than the actual activation of the unit.

For present purposes T, a parameter called temperature, is taken to be a fixed parameter that simply scales the magnitude of the net input. As before, trials are run by presenting input plus context, leaving them on until equilibrium is reached; and then selecting, at some particular time t, the most active member of the alternative set as the network's response. There is one minor complication: In Boltzmann machines, units representing particular alternatives are either on or off; therefore when a choice must be made, a tie is possible. In case of a tie, it is necessary to resample at a later time, repeating until no tie occurs.

The Independence Constraint on Architecture

When the above choice procedure is used, it can be shown that Boltzmann machines (symmetric, stochastic networks using the above binary update function) will conform to Morton's law as long as the network adheres to a particular architectural constraint. This constraint can be stated as follows:

1. The network must be partitionable into three sets of units: one set representing the alternatives, one set representing the stimulus input to the alternatives, and one set representing the context relevant to the alternatives.

2. No connections are allowed between any of the units in the input set and any of the units in the context set. The only interactions among the sets occurs by way of the units representing the alternatives.

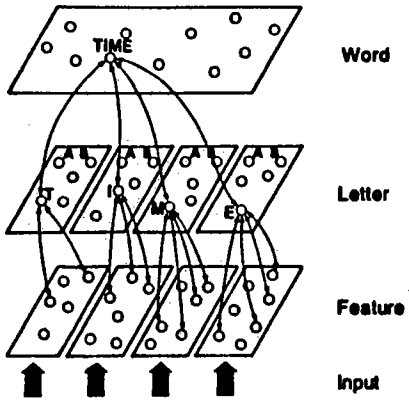

Figure 27.8 Architecture of the interactive activation model. Only selected connections are shown. All units within each enclosed region are mutually inhibitory. Reprinted from McClelland (1985).

The original interactive activation model, as formulated by McClelland and Rumelhart (1981) adhered to this architectural constraint, as shown in figure 27.8 (from McClelland 1985). For example, when the task is to identify the letter in a particular position (say, the second) the letter units in this position become the set representing the alternatives. The input set is the set of feature units for the second position, and the context set is all of the rest of the network, including the feature and letter units for other positions and including the word level units. There are no connections between any of the units in the input set and any of the context units, so the condition holds. Note that the condition holds simultaneously for all four letter positions. The crucial thing about this architectural constraint is that it preserves a kind of independence of the two sources of input. Neither source can influence the other—except by way of the units representing the alternatives. Therefore we will henceforth refer to this constraint as the independence constraint on network architecture, and we will refer to sources of input that are independent in this way as structurally independent.

To summarize, the analysis presented in McClelland (1991) shows that binary stochastic interactive networks that adhere to the independence constraint on architecture also adhere to the independence law of the effect of context on perceptual identification. Simulations suggest that the result holds for graded as well as binary activations.

Some readers may view the result of the analysis presented above as a minor matter, redressing an empirical inadequacy of the interactive activation model. But such a view misses the fact that the analysis actually represents an important advance in our understanding of the possible mechanistic basis of the independence law. This result shows that adherence to the independence law is an inherent characteristic of stochastic interactive networks that respect the architectural constraint. We need not conclude, as Massaro (1989) suggests

we should, that data adhering to Morton's independence law requires a strictly feedforward processing system; rather we now can see that an interactive mechanism with the right architecture is consistent with—even bound to produce—independence.

Using the Independence Constraint to Test Claims about Network Architecture

This result also leads to empirical tests that can distinguish between alternative possible network architectures. The independence law should hold, not only between stimulus and context, but more generally between any two sources of input that are structurally independent. Given this, suppose we manipulate two inputs that affect choices among a set of alternatives. If the independence law does not hold, this would establish that these inputs are not structurally independent. In a more agnostic vein, we can just run experiments, manipulating different inputs, and see whether the independence law holds. Note that cases in which the independence law does hold cannot strictly be used to argue for strict structural independence. There may well be structurally dependent architectures that produce close approximations to independence.

The method is analogous to the additive-factors method of Sternberg (1969a,b). According to this method, one manipulates two experimental factors. If they have non-additive effects on mean reaction time, we conclude that they exert their effects on one or more stages of processing in common. In what we will call the structural independence method, one manipulates two sources of input that influence stimulus identification response probabilities. If they have nonindependent effects on the z-transform of the probability of choosing a particular response, then we can conclude that the influences each source exerts on activations of response alternatives are not structurally independent.

If the architecture postulated for the interactive activation model is correct, we may find a failure of independence for effects of two different context letters on the identification of a target letter in a word. In the interactive activation model, the individual letters in the context of a target letter are not structurally independent. Rather, the individual context letters exert their influence on the identification of a target letter by way of a unit at the word level (see fig. 27.8). The nonlinearities in the network mean that they need not in this case have independent effects. This makes it possible for the effects of joint manipulations of context letters to violate independence.

Consider, for concreteness, the choice between the alternatives *I* and *O* for the identity of the middle letter in the following contexts: *L_E W_G L_G W_E*. Here, the lexical constraints are such that *LIE* and *WIG* are words, but *LOE* and *WOG* are not; while *LOG* and *WOE* are words but *LIG* and *WIE* are not. If we look at any one context letter (say, *L*), the constraint it places on the identity of the target letter reverses as we manipulate the identity of the other letter in the context. When the last letter is *E*, initial *L* favors middle *I*; but when the last letter is *G*, initial *L* favors middle *O*. Thus the constraint imposed

by one letter in the context is clearly dependent on the identity of the other letter. Now, if each context letter really does exert an independent effect on the identification of the target letter, then we expect no difference in the tendency to choose *I* in the contexts *W_G* and *L_E* compared to the contexts *L_G* and *W_E*. In other words, we expect no advantage for cases in which the target letter makes a word with the context, compared to cases in which it does not. On the other hand, if, as in the interactive activation model, each letter (target or context) influences activations of word units, and these in turn feed activation back to the letter level, then particular combinations of context letters may have differential effects. A word advantage, then, would support the interactive activation model's claim that the influences of different context letters on target identification are not structurally independent. Movellan and McClelland (n.d.) carried out an experiment to test for nonindependence in this case. Their subjects viewed brief, masked presentations of three-letter strings, and were given a forced-choice between two alternatives for one of the letters. In each case, neither context letter alone favored one alternative over the other, as in the above example; but the two context letters together always formed a word with only one of the two alternatives. If the two context letters were exerting independent effects on the probability of target identification, we should not expect a word advantage; if, however, the effects of the two context letters were not independent, but actually influenced target identification by way of their conjunction at the word level, then a word superiority effect should be obtained. The results confirmed the prediction of the interactive activation model: an accuracy advantage for words relative to nonwords was in fact obtained. The effect was relatively small, but was statistically reliable at the .001 level over both items and subjects.

There are other network architectures besides the one embodied in the interactive activation model. The described experiment does not rule out all alternatives to the architecture of the interactive activation model. But it does place constraints on the architecture, just as interactive effects of two factors on reaction times placed constraints on discrete stage models in Sternberg (1969a,b).

Summary

The present section has described how independent effects of context and stimulus information can arise in a stochastic interactive processing system and has demonstrated that independence need not always hold between different sources of information relevant to a particular identification response. Much remains to be done, of course. One task is to return to the kinds of situations that motivated the interactive activation model in the first place: situations in which context actually facilitates the discrimination of alternative letters (Reicher 1969). Such situations typically involve very brief presentations followed by a masking pattern—conditions in which activations clearly do not reach asymptote. One of the key findings from this paradigm is the discovery that the relative timing of context and stimulus information has a

big impact on performance; context must precede or be contemporaneous with the target for facilitation to occur. This and other aspects of the findings from this paradigm were accounted for by the original interactive activation model, and served to substantiate the claim that the phenomena required an account in terms of the detailed time course of information processing. For the present, though, the section has done three things relevant to the goals of the theory. First, it has illustrated how the principles described at the start are consistent with the effects of context and stimulus information on asymptotic choice responses; it has shown that the simple independence law can be an emergent result of processing systems that adhere to these principles; and it has made a first step toward using this result to begin establishing when independence will hold, and when it will not hold.

27.3.2 DYNAMICS OF PROCESSING: MONOTONIC AND NONMONOTONIC TIME-ACCURACY CURVES

This section will present some initial findings on the time-accuracy characteristics of information processing systems that are consistent with the principles of the GRAIN model. The theory is not as well developed for this case as for asymptotic activation; indeed it is generally easier to characterize the equilibrium states of nonlinear systems than it is to understand their time-dependent properties. In the absence of formal mathematical results, it is still possible to get some preliminary understanding of the variation of accuracy as a function of time in networks that adhere to the principles of the GRAIN model.

Wickelgren's Law

As in the previous section, we will consider a simple general regularity that has often been accounted for by simpler models, namely the fact that time-accuracy curves exhibit the shifted exponential form described by Wickelgren's law. Time-accuracy curves with approximately the right form arise naturally from models such as Ratcliff's diffusion model, in which there is a single continuous stochastic process. Such curves also arise when several gradual processes are organized in a feedforward cascade (McClelland 1979). Here we consider the case of multilayer graded, stochastic, processing systems with bidirectional excitatory connections between layers and bidirectional inhibitory connections within each layer.

The first point to note is the importance of distinguishing between propagation of activation and propagation of information in GRAIN networks. Let us think of the information propagation issue as follows. An experimenter selects one of *n* stimuli, and the subject's task is to indicate which it was through one of *n* corresponding responses. The stimulus is presented to the input layer of a multilayer GRAIN net, and begins a stochastic process of activation and inhibition that proceeds forward in time. For the sake of simplicity, we will consider a particular case in which there is one unit at each level corresponding

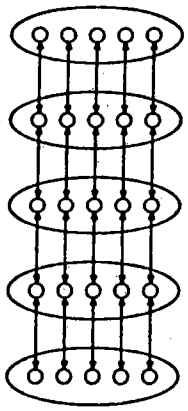

Figure 27.9 A simple five-layer GRAIN network used for exploring accumulation of information and RT distribution properties of GRAIN networks. Lines indicate bidirectional connections between levels; all units within each oval are mutually inhibitory.

to each stimulus, and we will assume that stimulus identification responses might be generated from any level at any time by selecting the most active unit at that level as the response. The propagation of information then consists of a gradual change in the sensitivity (d') of the response choice at each successive level to the actual identity of the stimulus presented as input. Simulations of d' as a function of time were carried out using the network of figure 27.9. In this network, there are five levels, each with five units. Units at each level have bidirectional excitatory connections with corresponding units at the next level. Within each level, there are bidirectional inhibitory connections among all the units. The details of the simulation follow, since they are not reported elsewhere. Each trial began with all activations set to 0, and was followed by 200 initialization cycles, with no external input. Stimulus onset then occurred. This consisted of applying external input of 3.0 to the middle unit at the input level. Processing continued for 2000 cycles. On each cycle, net input to each unit was first computed, then all activations were updated. The net input consisted of a fixed negative bias of -3.0, external input if any, a sample of Gaussian noise with mean 0 and standard deviation 1.0, and the summed input from all other units. Between level excitatory weights were 5.0; within level inhibitory weights were -1.5. The activation of each unit was then incremented by an amount equal to k_l(logistic(net) $-$ act), where k_l represents a rate constant for all the units at a particular level, net represents the net input and act represents the activation of the unit from the previous time step.

In calculating d', we need to know the hit rate, or the probability that the correct alternative is the most active at each level at each time step. This is calculated simply by keeping a tally over many repeated trials. We also need a false alarm rate, which would ordinarily be the probability that the middle alternative is active when some other input is presented. Due to the fact that parameters are equivalent for all alternatives, the probability that the middle alternative would be most active when some other alternative is shown is equal to the probability that some other alternative is most active when the middle alternative is shown. So the false alarm rate was estimated from this latter probability.

The results of two simulation runs of 5000 trials each using the network of figure 27.9 are shown in figure 27.10. What we see in each case is that, even though activation flows both ways in this network, information in the sense of sensitivity to the stimulus propagates from the input forward. Each successive level exhibits a shifting and slowing of the growth of d', in approximate conformity to the cascade model. When all the levels have the same rate constant, the cascade model produces time-accuracy curves that converge to a sigmoid as more and more levels are added. The same thing happens in the GRAIN network, as shown in the upper panel, where the rate constant, k_l is set to 0.15 for all levels.

There are some differences between the behavior of GRAIN networks and the behavior of the cascade model. In the cascade model, when the rate constant of one level is slowed, it affects the time course of processing at that level and all subsequent levels. In the case where one level has a slower rate constant than all the others, the time-accuracy curves generated by the cascade model have a clear shifted exponential shape, with the slowest rate constant determining the rate constant of the curve, and the other rate constants determining merely the shift of the takeoff point (McClelland 1979). In the present GRAIN network, on the other hand, when the rate constant of one level is slowed relative to the rate constants of the others, the effect spreads throughout the network, and the results are difficult to distinguish from the case in which all of the levels have the same rate constants. This is seen in the lower panel of figure 27.10, where the rate constant of the third level is set to the slow value of 0.05, while the rate constants of all of the other levels are set to 0.30. In a cascade model these values would produce a very clearly marked change in the rate of rise of the third curve; the fourth and fifth curves would look almost exactly like the third, but just successively shifted over.

This simulation shows, then, that the conformity of the GRAIN model to Wickelgren's law is somewhat less exact than its conformity to Morton's law. The time-accuracy curves are not strictly or even closely shifted exponentials but are perhaps better described as slightly skewed sigmoids. Even so, the curves do obey the general qualitative form of the empirical findings. Accuracy starts at chance and, after a flat period, makes a transition to a relatively steep, then negatively accelerating, approach to asymptote. Whether the exact form of empirical time-accuracy curves is consistent with this pattern is not completely clear at this time. Furthermore, it seems quite possible that interactivity may not hold between perceptual and response selection processes. In that case, if response selection is a rate-limiting process, the overall resulting time-accuracy curve would look indistinguishable from the curves generated by the cascade model and would closely approximate Wickelgren's law.

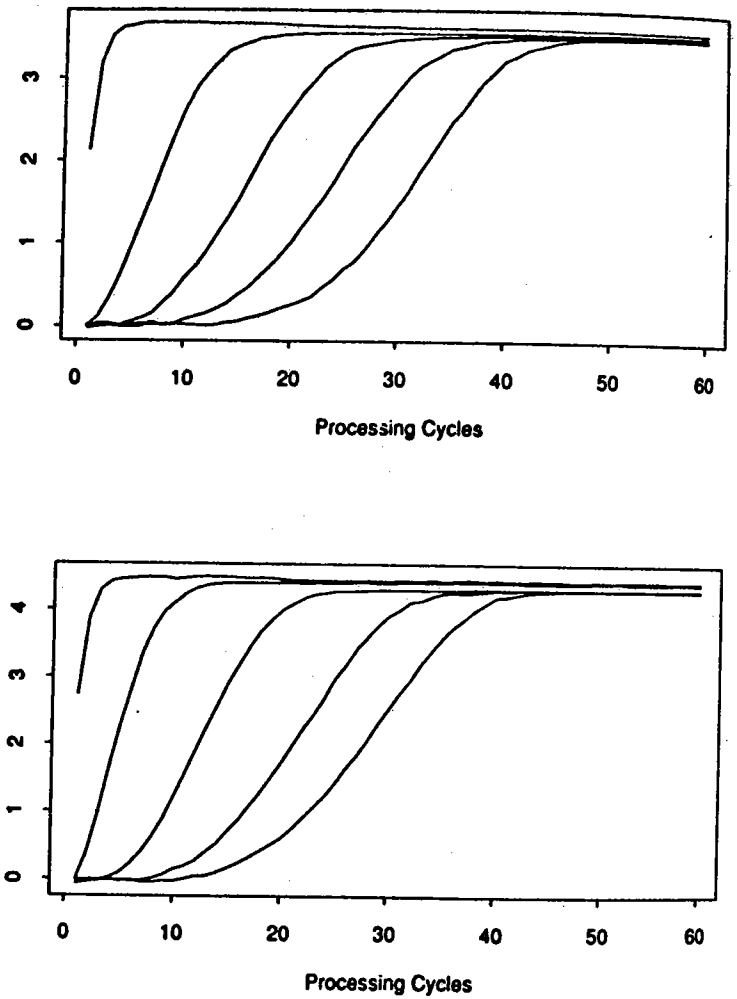

Figure 27.10 The buildup of sensitivity to the input at successive levels in the GRAIN network shown in figure 27.9, for four different values of the parameter lambda. Note that the same value of lambda applies to all of the units in the network in these simulations. Each panel shows d' at the first, second, third, fourth, and fifth level of the processing system shown in figure 27.9. In all cases the curves for successively deeper levels arrange themselves from left to right.

While GRAIN models can conform to the general form of Wickelgren's law, there are circumstances in which they will not. One factor that appears to be important is resting activation level, as determined in this case by the bias in the net input. A large negative bias is required, so that resting activations tend to be quite low, or else the network will exhibit a tendency to become relatively committed to a spurious pattern of activation prior to stimulus presentation. In such cases, information can propagate extremely slowly through the system, and the time-accuracy curves have a rather different shape.

Even when parameters are such that information propagates promptly through the system, a Wickelgren-like shape to time-accuracy curves is not strictly ensured; indeed, the curve need not even by monotonic, as will be seen below. A statement of the conditions under which approximate conformity to an exponential approach to asymptote can be obtained is beyond the reach of the present analysis. However it is possible to offer one observation concerning conditions that may lead to violations of monotonicity. The observation is simply that time-accuracy curves may fail to be monotonic when the ordering of expected values of asymptotic activations of units would be different without lateral inhibition than with lateral inhibition. In the simulation reported below, lateral inhibition does change the ordering of expected values of asymptotic activations of response units in one experimental condition, and a U-shaped time-accuracy curve is obtained.

A U-shaped Time-Accuracy Curve

It is one thing to know that a process model adhering to a set of principles can account for data that exhibit a general regularity. This state of affairs allows us to suppose that the principles in question may in fact be part of a useful description of the underlying processes. But the question arises, what does the process model buy us, if we already have a general law that we can refer to that correctly characterizes the outcomes of experiments? One answer to this question is that it may allow us to give straightforward accounts for results obtained in cases where existing data actually violates the general regularities.

Such a case is provided by the interesting experiment of Gratton et al. (1988). In this experiment, subjects viewed target letters flanked by two letters on each side. The task was to indicate whether the target was *S* or *H*. The two targets (*S* and *H*) could be flanked either by the congruent letter (as in *SSSSS* and *HHHHH*) or by the other letter (*SSHSS* and *HHSHH*). Subjects were induced to respond so quickly that they produced about 15 percent errors over all. When Gratton et al. looked at accuracy in the two conditions, conditional on response time, they found a Wickelgren-like conditional accuracy function in the congruent condition, but a U-shaped curve in the incongruent condition. The results are shown in figure 27.11 (top). The conditional accuracy functions are shown along with the distribution of response times by bin, for each condition. (Actually the graph shows time and accuracy for electromyographically detected muscle activity that precedes the response. The actual RT data

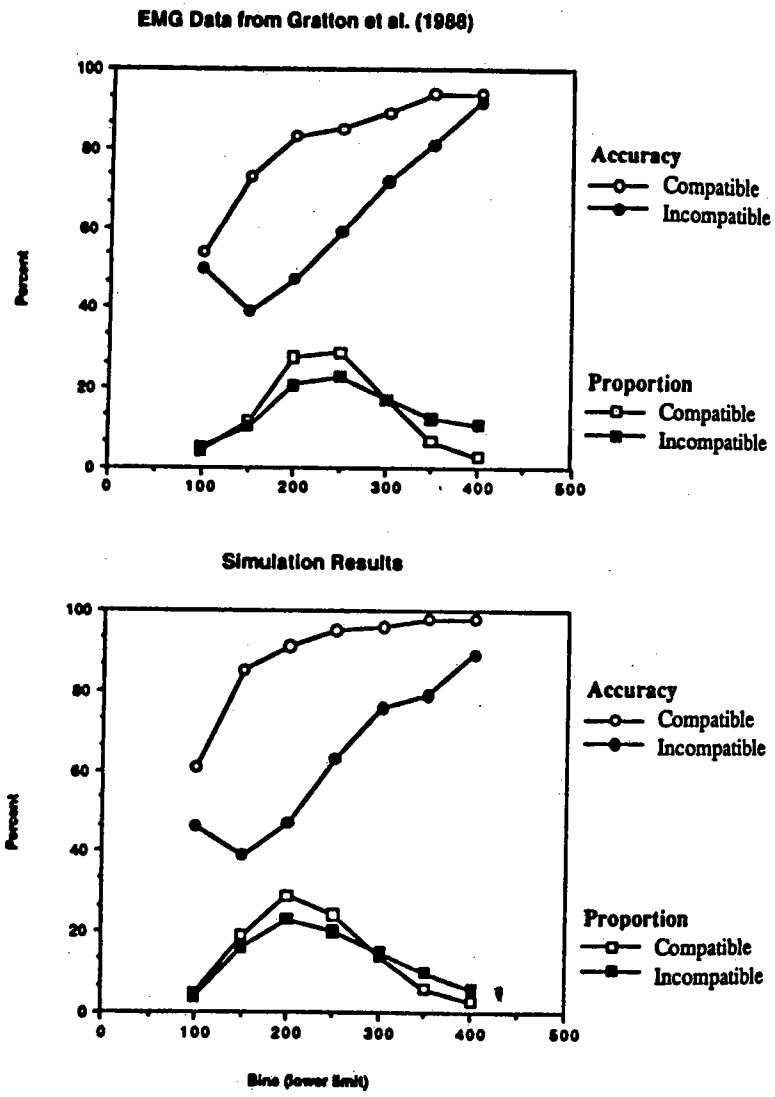

Figure 27.11 Data and simulation of Gratton et al. (1988). RTs are actually time from stimulus onset to the detection of electromyogram (EMG) activity in the muscles innervating the executed response, though actual RTs show the same effects. The top panel represents the experimental data; the bottom panel represents the results of the simulation. Within each panel, the top two curves represent accuracy in each time interval; the bottom two curves represent the proportion of trials falling in each bin. Reprinted with permission from Cohen, Servan-Schreiber, and McClelland (1992).

looks much the same, except that all times are somewhat longer). Gratton et al. interpreted their results in the following way. They suggested that there are two letter identification pathways, one, which is fast, for identification of letters without regard to position; and another, slower one, for position-specific identification. Their idea was that the fast pathway will report the predominant identity present in the display, and this will lead to the dip if a response is initiated before target-specific information arrives from the position-specific pathway.

The GRAIN framework provides us with an opportunity to propose an alternative interpretation. In accord with the observation made at the end of the preceding section, it can be proposed that the U-shaped curve reflects the resolution of a competition process driven by mutual inhibition. Servan-Schreiber (1990) has developed a GRAIN model that captures this notion and provides a nice account of the pattern of data obtained in the experiment.

The model begins with the idea, common to many recent connectionist models of spatial attention (e.g., Cohen, Dunbar, and McClelland 1990; Phaf, van der Heiden, and Hudson 1990; Mozer 1991; LaBerge and Brown 1989) that attention serves to prime feature detectors for attended locations. This gives the detectors for the attended location a very slight initial advantage over detectors in other locations. Early in processing, features in the target location have only a slight advantage over features in non-target locations, allowing them to conspire against the target in incongruent displays. Mutual competition then causes this initial advantage to become accentuated as the detectors become activated by the stimulus.

The network Servan-Schreiber used to capture these ideas consists of three pools of units: A pool consisting of position-specific feature analyzer units (for features of letters in each of three positions), an output layer consisting of a response unit for each alternative, and an attention layer consisting of position-specific attention-to-location units (see fig. 27.12). While Gratton et al. used a target with two flankers on each side, the flanker effect can be captured easily enough with only one, for a total of three display positions. So there are only three positional attentional units (for left, center, and right locations). In the feature pool, there are feature analyzers for features of *S* and *H* in each position; for simplicity only one *S*-feature unit and one *H*-feature unit is included for each position. Within each pool, units are mutually inhibitory. Between pools, there are mutually excitatory connections between *S* features and the *S* output unit, and between *H* features and the *H* output unit. There are also mutually excitatory connections between the attentional unit for a particular position and the units for features in the corresponding position. A response is recorded as soon as one of the output units reaches a fixed threshold.

Consider, now, what happens when a display is presented containing an *H* flanked by an *S* on each side. Attention to the center position is implemented by external input to the position-specific attention unit for the target location. This input is turned on and the network is allowed to stabilize before the trial begins. During stabilization, some activation spreads to the corresponding

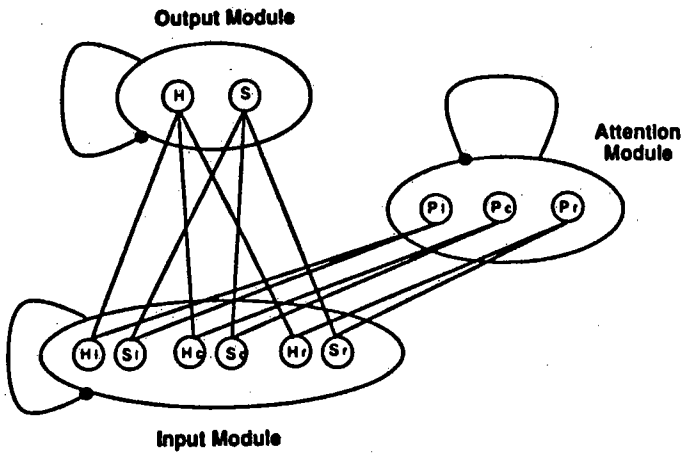

Figure 27.12 Network used to simulate the U-shaped time-accuracy curve of Gratton et al. (1988).

position-specific feature units, giving the detector for the features in the target location a very slight advantage over features in other positions. This slight advantage is maintained when the target is presented. However, since there are two nontargets but only one target, the aggregate activation for *S* features is almost twice as great as the aggregate activation for *H* features. Without mutual competition, this state of affairs would persist indefinitely in this network. But with mutual competition, the target features gradually gain the upper hand over the features of the nontargets. This happens because the features of each letter are in competition with the features of each other letter, and because of the slight initial advantage for the target features due to top-down activation from attention. The ability of items with a slight initial advantage to win out over a number of competing alternatives was examined by McClelland and Rumelhart (1981); in that model it allowed words with a very small advantage in resting activation level (due to higher frequency) to win out over other words of lower frequency.

The simulation results are shown below the data from Gratton et al. in figure 27.11. The simulation of the compatible condition follows the conventional gradual approach to asymptote, as found by Gratton et al. In the incompatible condition, the curve follows the U-shaped pattern seen in the empirical data. This occurs, not because of separate pathways for position-dependent and position-independent information, but because of an interactive activation and competition process that initially favors the flankers due to their greater number but eventually favors the letter in the attended position.

The simulation also captures several other aspects of the Gratton et al. data. One of these is the general shape of the reaction-time distributions found in each condition of the experiment (see fig. 27.11), and the effect of compatibility on these distributions. Another is a fact that emerged from the analysis by

Gratton et al. of ERP (event-related potential) data recorded from subjects in their experiment. They recorded ERPs over the motor area contralateral to each responding hand as well as EMGs in the muscles governing each response, and found that responses appeared to be triggered when the difference in the evoked response over the two motor cortices reached a particular threshold value. Even though the simulations of Servan-Schreiber (1991) actually used a fixed threshold, rather than a difference signal, to trigger the response, it turned out that the difference in activation of the two response units at the time the response was triggered exhibited the same constant difference that Gratton et al. found in their ERP data.

Let us now consider how the various principles of the GRAIN model contribute to the account for the data. Competitive inhibition at the position-specific feature level plays a crucial role in producing the U-shaped form of the time-accuracy curve in the incompatible condition, as already explained. A feature of competitive inhibition is that it only comes into effect after some initial activation; it is this initial activation that actually gives rise to the dip in the curve, and the competitive inhibition that ultimately results in the correct response dominating performance, correcting the dip. Competitive inhibition at the response level plays a role in another aspect of the model, as well. It causes a fixed activation threshold to actually correspond to a fixed difference in activation between the two alternatives. Both of these effects of competitive inhibition are, however, dependent on graded and gradual propagation of activation. Graded and gradual propagation allow the summed influences of the flankers to initially govern the preferred response, while also allowing the balance to swing gradually back in favor of the correct alternative as processing continues. Indeed, Coles (1989) argues that aspects of the findings of Gratton et al. actually demonstrate the propagation of graded information.

Intrinsic variability also plays an important enabling role. This variability has the effect of causing variability in the activation of response units at stimulus onset, and in the subsequent activation process. (Variability at time of onset may be crucial in the present simulations, since it alone would be sufficient to produce a broad distribution of reaction times.) The U-shaped time-accuracy curves arise primarily from trials in which the incorrect alternative happens to be relatively more active at the time of stimulus onset so that the slight early activation advantage of the incorrect alternative pushes the incorrect alternative over the threshold.

We have seen, then, that the model relies on competition, graded and gradual propagation of activation, and on variability. On the other hand, interactivity does not appear to contribute importantly to the account of the data. Servan-Schreiber (verbal communication, 1991) has found that in fact the data can be fit quite well in a network with only unidirectional connections between levels (input and position specific locations to position-specific features, position-specific features to responses). Interactivity is perfectly compatible with the findings, and there are many reasons to favor an interactive account of attention and perception (Cohen et al. in press), but this particular feature of GRAIN models appears not to be crucial in this case.

Summary

This second case study indicates, once again, some progress on each of the three goals of theory development. First, we have seen how the principles of GRAIN can begin to offer an account for aspects of data obtained within the time-accuracy paradigm. Second, we have seen how relatively complex processing systems (i.e. with as many as five levels) can in some cases give rise to time-accuracy curves that conform approximately to a simple general law, in this case Wickelgren's law. Correspondingly we have seen in another case how GRAIN provides a framework for understanding one particular case in which this simple general law breaks down. In the case in point, GRAIN appears to provide a fairly natural and direct account, without requiring that responses be based at different times on the output of different processing systems. Third, some small steps have been taken toward an understanding of which of the principles are responsible for the simulation of the experimental results, and which are merely compatible with them.

While these results are suggestive, it should be clear that we are far from a full account of the relation between GRAIN networks and time-accuracy curves. It is far from clear, for example whether the forms of typical empirical time-accuracy curves are really more accurately described by Wickelgren's shifted exponentials than by GRAIN's skewed sigmoids. Nor is it clear in any detail what conditions will lead even to monotonicity, and what conditions will produce U-shaped time-accuracy curves. These issues are among those that must be addressed, if the theory is to be advanced beyond its present suggestive but preliminary state.

27.4 DISCUSSION

Obviously, the initial steps taken here leave us far from a complete theory of the asymptotics and dynamics of information processing. Of the three information-processing paradigms that the theory is intended to unify, only two (the asymptotic accuracy paradigm and the time-accuracy paradigm) have really been considered at all. Analytic results are available only for asymptotic accuracy. Nevertheless we can begin to see three contributions emerging from these preliminary explorations.

First, this research has already begun to help us to see how general regularities of human performance data might arise as emergent properties of networks of simple computing elements that adhere to the principles of GRAIN. Morton's independence law and Wickelgren's characterization of time-accuracy curves are cases in point. As such, the work is a first step toward understanding how the mass of processing activity triggered by the presentation of a stimulus gives rise to the simple and regular outcomes that are typically observed in behavior.

Second, the work begins to give us ways of predicting violations of these simple regularities (in the case of the proposed test of nonindependent effects of context letters), and of providing a basis for accounting for observed deviations from typical outcomes (as in the case of the Gratton et al. U-shaped time-accuracy curves). Thus the theory allows a significant advance in our understanding beyond the simple statement of the regularities themselves.

Third, this research begins to provide a unification of the disparate literatures on the dynamics of information processing on the one hand and of the effects of context on asymptotic choice performance on the other. In this regard it may be worth noting that such a unification seems to be possible only because a common set of principles is used in both cases. Many models of asymptotic choice behavior use time-independent functions of graded variables (e.g., strengths), while many models of the time course of processing use time-dependent functions of discrete variables (e.g., subprocess X is or is not complete by time t). Of course, the present framework is not the first to consider the propagation of graded information; the point here is only that this may well be a necessary, though surely not a sufficient, condition for unifying the analysis of the dynamics and the asymptotics of information processing.

Next Steps for Experimental Research

If the unification of these disparate literatures is to be complete, and if the exploration of the detailed dynamics of processing is to be taken further, it will become necessary to conduct a greater number of studies exploring the time course of processing. Up to now relatively few such studies have been produced. One reason for this has been the apparent difficulty of collecting time-accuracy data. It may be, however, that the difficulty of time-accurracy studies is more apparent than real. In the author's laboratory we have recently been able to replicate the Gratton et al. study, complete with a dip below chance in the incongruent condition, in an experiment in which each of ten unpracticed subjects came into the lab and performed in 800 trials (400 per condition) in less than 45 minutes. The method is very simple. The subject is instructed to go as fast as possible, and is rewarded for fast correct responses (a penny for each RT less than 300 ms). RTs range from less than 100 to more than 500 ms, with the bulk between 150 and 400, which spans the dip in the Gratton et al. curve.

Given the simplicity of the method, it seems likely that it is feasible to trace more thoroughly than has typically been done in the past the relation between time since stimulus onset and response accuracy. If so, this should help us to see whether the GRAIN model, some refinement of it, or some completely different kind of account, ends up providing the best characterization of the time course and the outcome of information processing.

NOTES

The work reported here was supported by a NIMH Career Development Award MH-00385. Thanks for valuable comments on earlier drafts of this article are due to the editors, to two anonymous reviewers, and to Jonathan Cohen, Javier Movellan, and Leigh Nystrom. Address

correspondence concerning this manuscript to the author at Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA.

1. There is now a considerable body of work on learning rules that operate by making symmetric adjustments to training weights. Such networks may be initialized with non-symmetric weights, which are naturally symmetrized in the course of training (Hinton 1989; see Hertz, Krogh, and Palmer 1990 for a discussion of this issue). In practice it is generally found that it is not even strictly necessary for every unit to be connected to every other unit; networks that use distributed representations come to behave as though they were fully symmetric, even when the symmetry is not maintained on a unit-by-unit level.

2. Technically, this property holds only when the network is both Markovian (so that the next state depends only on the preceding state) and ergodic (so that every state can be reached from every other state.) GRAIN networks have both of these properties.

REFERENCES

Anderson, James A. (1991). Why, having so many neurons, do we have so few thoughts? In W. E. Hockley and S. Lewandowsky (Eds.), *Relating theory and data*. Hillsdale, NJ: Erlbaum.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361

Cohen, J. D., Servan-Schreiber, D., and McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *American Journal of Psychology, 105,* 239–269.

Coles, Michael G. H. (1989). Modern mind-brain reading: psychophysiology, physiology, and cognition. *Psychophysiology, 26,* 251–269.

Cox, D. R., and Miller, H. D. (1965). *The theory of stochastic processes*. New York: Wiley

Elman, J. L., and McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165.

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

Gratton, G., Coles, M. H., Sirevaag, E. J., Eriksen, C. W., and Donchin, E. (1988). Pre- and poststimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 331–344.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part 1. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134.

Grossberg, S. (1978a). A theory of human memory: Self-organization and performance of sensory motor codes, maps and plans. In R. Rosen and F. Snell (Eds.), *Progress in theoretical biology*. Vol. 5, 233–374. New York: Academic Press.

Grossberg, S. (1978b). A theory of visual coding, memory, and development. In E. L. J. Leeuwenberg and H. F. J. M. Buffart (Eds.), *Formal theories of visual perception*. New York: Wiley.

Hertz, J. Krogh, A., and Palmer, R. (1990). *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation, 1,* 143–150.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland and the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA: MIT Press.

Hinton, G. E., and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland and the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1, *The PDP perspective*. Cambridge, MA: MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554–2558.

LaBerge, D., and Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review, 96,* 101–124.

Lappin J. S. and Disch, K. (1973). The latency operating characteristic: 11. Effects of visual stimulus intensity on choice reaction time. *Journal of Experimental Psychology, 93,* 367–372.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science, 9,* 113–146.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L., and O'Regan, J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 634–644.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, Part I: An account of basic findings. *Psychological Review, 88,* 375–407.

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

McClelland, J. L., and Rumelhart, D. E. (1988). *Explorations in parallel distributed processing. A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.

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

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21,* 398–421.

Minsky, M., and Pappert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Morton J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Movellan, J., and McClelland, J. L. (1991). Learning continuous probability distributions with the contrastive Hebbian algorithm. *Technical Report PDP-CNS-92-2,* Department of Psychology, Carnegie-Mellon University.

Movellan, J., and McClelland, J. L. (N.d.). Theoretical and empirical consequences of the stochastic interactive activation model. Unpublished.

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach*. Cambridge, MA: MIT Press.

Newell, A. (1973) . You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Newell, A. (1990). Metaphor for mind, theories of mind, should the humanities mind? In J. Sheehan and M. Sosna (Eds.), *Boundaries of humanity: Humans, animals and machines*. Berkeley, CA: University of California Press.

Oden, G. C., and Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85*, 172–191.

Pachella, R. (1974). The interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. New York: Halstead.

Phaf, R. H., van der Heijden, A. H. C., and Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*, 273–341.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 274–280.

Rumelhart, D. E. (1977). Understanding and summarizing brief stories. In D. LaBerge and S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*, 265–303. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, and the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition*. Vol. 1, *Biological mechanisms*. Cambridge, MA: MIT Press.

Rumelhart, D. E., and McClelland, J. L. (1981). Interactive processing through spreading activation. In C. Perfetti and A. Lesgold (Eds.), *Interactive processes in reading*. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception. Part II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*, 60–94.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart, J. L. McClelland, and the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1, *Foundations*. Cambridge, MA: MIT Press.

Servan-Schreiber, D. (1990). From physiology to behavior: Computational models of catecholamine modulation of information processing (Ph.D. thesis). *Technical Report CMU-CS-90-167*, School of Computer Science, Carnegie-Mellon University.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In D. E. Rumelhart, J. L. McClelland, and the PDP research group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognitron*. Vol. 2, Psychological and Biological models. Cambridge, MA: MIT Press.

Sternberg, S. (1969a). The discovery of processing stages: Extensions of Donders' method. In W. G. Koster (Ed.), *Attention and performance*. Vol. 2. Amsterdam: North-Holland.

Sternberg, S. (1969b). Memory-scanning: Mental processes revealed by reaction time experiments. *American Scientist, 57*, 421–457.

Wickelgren, W. A. (1977). Speed accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67–85.

Williams, R. J. (1986). The logic of activation functions. In D. E. Rumelhart, J. L. McClelland, and the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1, *Foundations*. Cambridge, MA: MIT Press.