**Neural Representation and Neural Computation**

Patricia Smith Churchland; Terrence J. Sejnowski

*Philosophical Perspectives*, Vol. 4, Action Theory and Philosophy of Mind. (1990), pp. 343-382.

Stable URL:
http://links.jstor.org/sici?sici=1520-8583%281990%294%3C343%3ANRANC%3E2.0.CO%3B2-P

*Philosophical Perspectives* is currently published by Blackwell Publishing.

---

Philosophical Perspectives, 4
Action Theory and Philosophy of Mind, 1990

# NEURAL REPRESENTATION AND NEURAL COMPUTATION

Patricia Smith Churchland
University of California at San Diego

Terrence J. Sejnowski
Johns Hopkins University

The types of representation and the styles of computation in the brain appear to be very different from the symbolic expressions and logical inferences that are used in sentence-logic models of cognition. In this chapter we explore the consequences that brain-style processing may have on theories of cognition. Connectionist models are used as examples to illustrate neural representation and computation in the pronouncing of English text, and in the extracting of shape parameters from shaded images. Levels of analysis are not independent in connectionist models, and the dependencies between levels provide an opportunity to co-evolve theories at all levels. This is a radical departure from the *a priori*, introspection-based strategy that has characterized most previous work in epistemology.

## 1. How Do We Represent the World?

The central epistemological question, from Plato on is this: *How is representation of a world by a self possible?* So far as we can tell, there is a reality existing external to ourselves, and it appears that we do come to represent that reality, and sometimes even to know that its initial appearance to our senses differs from how it actually is. How is this accomplished, and how is knowledge possible? How is science itself possible?

The dominant philosophical tradition has been to try to resolve the epistemological puzzles by invoking mainly intuition and logic

to figure out such things as the organization of knowledge, the nature of the "mirroring" of the outer world by the inner world, and roles of reason and inference in the generation of internal models of reality. Epistemology thus pursued was the product of "pure reason", not empirical investigation, and thus epistemological theories were believed to delimit the necessary conditions, the absolute foundations, and the incontrovertible presuppositions of human knowledge. For this *a priori* task—a task of reflective understanding and pure reason—empirical observations by psychologists and neurobiologists are typically considered irrelevant, or at least, incapable of effecting any significant correction of the *a priori* conclusions. Plato, Descartes, and Kant are some of the major historical figures in that tradition; some contemporary figures are Chisholm (1966), Strawson (1966), Davidson (1974), and McGinn (1982). It is safe to say that most philosophers still espouse the *a priori* strategy to some nontrivial extent.

In a recent departure from this venerable tradition of *a priori* philosophy, some philosophers have argued that epistemology itself must be informed by the psychological and neurobiological data that bear upon how in fact we represent and model the world. First articulated in a systematic and powerful way by Quine (1960)[1], this new "naturalism" has begun to seem more in keeping with evolutionary and biological science and to promise more testable and less speculative answers.

If, as it seems, acquiring knowledge is an essentially biological phenomenon, in the straightforward sense that it is something our brains do, then there is no reason to expect that brains should have evolved to have *a priori* knowledge of the true nature of things: not of fire, not of light, not of the heart and blood, and certainly not of knowledge, or of its own microstructure and microfunction. There are, undoubtedly, innate dispositions to behave in certain ways and to believe certain things, and to organize data in certain ways, but innateness is no guarantee of truth, and it is the truth that *a priori* reflections are presumed to reveal. Innate beliefs and cognitive structures cannot be assumed to be either optimal or true, because all evolution "cares" about is that the internal models enable the species to survive. Satisficing is good enough. It is left for science to care about the truth (or perhaps empirical adequacy), and the theories science generates may well show the inadequacies of our innately specified models of external reality. Even more dramatically,

they may show the inadequacy of our model of our internal reality—of the nature of our selves.

The *a priori* insights of the Great Philosophers should be understood, therefore, not as The Absolute Truth about how the mind-brain must be, but as articulations of the *assumptions* that live deep in our collective *conception* of ourselves. As assumptions, however, they may be misconceived and empirically unsound, or at least they may be open to revision in the light of scientific progress. The possibility of such revision does not entail that the assumptions are ludicrous or useless. On the contrary, they may well be very important elements in the theoretical scaffolding as neurobiology and psychology inch their way toward empirically adequate theories of mind-brain function. The methodological point is that in science we cannot proceed with no theoretical framework, so even intuitive folk theory is better than nothing as the scientific enterprise gets underway.

In addition to asking how the self can know about the external reality, Kant asked: How is representation of a *self* by a knowing self possible? One of his important ideas was that the nature of the internal world of the self is no more unmediated or *given* than is knowledge of the external world of physical objects in space and time. A modern version of this insight says: just as the inner thoughts and experiences may represent but not *resemble* the outer reality, so the inner thoughts may represent but not resemble the inner reality of which they are the representation. This idea, taken with Quine's naturalism, implies that if we want to know how we represent the world—the external world of colored, moving objects, and the internal world of thoughts, consciousness, motives and dreams—the scientific approach is likely to be the most rewarding. Inner knowledge, like outer knowledge, is conceptually and theoretically mediated–it is the result of complex information processing. Whether our intuitive understanding of the nature of our inner world is at all adequate is an empirical question, not an *a priori*, one.

If empirical results are relevant to our understanding of how the mind-brain represents, it is also entirely possible that scientific progress on this frontier will be as revolutionary as it has been in astronomy, physics, chemistry, biology, and geology. With this observation comes the recognition that it may reconfigure our current assumptions about knowledge, consciousness, representations, and the self at least as much as Copernicus and Darwin reconfigured our

dearest assumptions about the nature of the universe and our place in it. Our intuitive assumptions, and even what seems phenomenologically obvious, may be misconceived and may thus undergo reconfiguration as new theory emerges from psychology and neurobiology.

Philosophers—and sometimes psychologists, and occasionally even neuroscientists—generally make one of two responses to the naturalists' conception of the status of our self-understanding:

(1) Philosophy is an *a priori* discipline, and the fundamental conceptual truths about the nature of the mind, of knowledge, of reason, etc. will come only from *a priori* investigations. In this way, philosophy sets the bounds for science—indeed, the bounds of sense, as Strawson (1966) would put it. In a more extreme vein, some existentialist philosophers would claim that the naturalistic approach is itself symptomatic of a civilizational neurosis: the infatuation with science. On this view, the scientific approach to human nature is deeply irrational. Mandt (1986, p. 274) describes the existentialist criticism as follows: "that scientific modes of thought have become paradigmatic indicates the degree to which traditional modes of human life and experience have disintegrated, plunging civilization into a nihilistic abyss."

(2) Even if a naturalistic approach is useful for some aspects of the nature of knowledge and representation, the neurosciences in particular are largely irrelevant to the enterprise. Neuroscience may be fascinating enough in its own right, but for a variety of reasons it is irrelevant to answering the questions we care about concerning cognition, representation, intelligent behavior, learning, consciousness, and so forth. Psychology and linguistics might actually be useful in informing us about such matters, but neurobiology is just off the book.

## 2. Why Is Neurobiology Dismissed as Irrelevant to Understanding How the Mind Works?

### *2.1 The Traditional Problem*

In its traditional guise, the mind-body problem can be stated thus: are mental phenomena (experiences, beliefs, desires. etc.) actually

phenomena of the physical brain? Dualists have answered No to this question. On the dualist's view, mental phenomena inhere in a special, nonphysical substance: the mind (also referred to as the soul or the spirit). The mind, on the dualist's theory, is the ghost in the machine; it is composed not of physical material obeying physical laws, but of soul-stuff, or "spooky" stuff, and it operates according to principles unique to spooky stuff.

The most renowned of the substance dualists are Plato and Descartes, and more recently, J. C. Eccles (1977) and Richard Swinburne (1986). Because dualists believe the mind to be a wholly separate kind of stuff or entity, they expect that it can be understood only in its own terms. At most, neuroscience can shed light on the *interaction* between mind and body, but not on the nature of the mind itself. Dualists consequently see psychology as essentially independent of neurobiology, which, after all, is devoted to finding out how the *physical* stuff of the nervous system works. It might be thought a bonus of dualism that it implies that to understand the mind we do not have to know much about the brain.

Materialism answers the mind-body question (are mental states actually states of the physical brain?) in the affirmative. The predominant arguments for materialism draw upon the spectacular failure of dualism to cohere with the rest of ongoing science. And as physics, molecular biology, evolutionary biology, and neuroscience have progressed, this failure has become more rather than less marked. In short, the weight of empirical evidence is against the existence of special soul-stuff (spooky stuff). (For a more thorough discussion of the failures of substance dualism, see P. S. Churchland 1986.) Proponents of materialism include Hobbes in the seventeenth century, and in the twentieth, B. F. Skinner (1957, 1976), J. J. C. Smart (1959), W. V. O. Quine (1960), D. C. Dennett (1978) and P. M. Churchland (1984).

Despite the general commitment to materialism, there are significant differences among materialists in addressing the central question of how best to explain psychological states. Strict behaviorists, such as Skinner, thought that explanations would take the form of stimulus-response profiles exclusively. Supporting this empirical hypothesis with a philosophical theory, philosophical behaviorists claimed that the mental terminology itself could be analyzed into sheerly physicalistic language about dispositions to behave. (For discussion, see P. M. Churchland 1988) Curiously,

perhaps, the behaviorists (both empirical and philosophical) share with the dualists the conviction that it is not necessary to understand the workings of the brain in order to explain intelligent behavior. On the behaviorists' research ideology, again we have a bonus: In order to explain behavior, **we do not have to know anything about the brain.**

In contrast to behaviorism, identity theorists (Smart 1959, Enc 1983) claimed that mental states, such as visual perceptions, pains, beliefs, and drives, were in fact identical to states of the brain, though it would of course be up to neuroscience to discover precisely what brain states were in fact identical to what mental states. On the research ideology advocated by these materialists, explanation of behavior will have to refer to inner representations and hence to what the brain is doing.

### *2.2. The Contemporary Problem: Theory Dualism*

Many philosophers who are materialists to the extent that they doubt the existence of soul-stuff nonetheless believe that psychology ought to be essentially autonomous from neuroscience, and that neuroscience will not contribute significantly to our understanding of perception, language use, thinking, problem solving, and (more generally) cognition. Thus, the mind-body problem in its contemporary guise is this: Can we get a *unified* science of the mind-brain? Will psychological theory reduce to neuroscience?

A widespread view (which we call Theory Dualism) answers No to the above question. Typically, three sorts of reasons are offered:

*Neuroscience is too hard.* The brain is too complex; there are too many neurons and too many connections, and it is a hopeless task to suppose we can ever understand complex higher functions in terms of the dynamics and organization of the neurons.

*The argument from multiple instantiability.* Psychological states are functional states and, as such, can be implemented (instantiated) in diverse machines (Putnam 1967, Fodor 1975, Pylyshyn 1984). Therefore, no particular psychological state, such as believing that the earth is round or that $2 + 2 = 4$, can be identified with exactly this or that machine state. So no functional (cognitive) process can be reduced to the behavior of particular neuronal systems.

*Psychological states have intentionality.* That is, they are identified in terms of their semantic content; they are "about" other things;

they represent things; they have logical relations to one another. We can think about objects in their absence, and even of nonexistent objects. For example, if someone has the belief that Mars is warmer than Venus, then that psychological state is specified as the state it is in terms of the sentence "Mars is warmer than Venus", which has a specific meaning (its content) and which is logically related to other sentences. It is a belief *about* Mars and Venus, but it is not caused by Mars or Venus. Someone might have this belief because he was told, or because he deduced it from other things he knew. In cognitive generalizations states are related semantically and logically, whereas in neurobiological generalizations states can only be *causally* related. Neurobiological explanations cannot be sensitive to the logical relations between the contents of cognitive states, or to meaning or "aboutness". They respond only to *causal* properties. Neurobiology, therefore, cannot do justice to cognition, and thus no reduction is possible.

### *2.3. What Is Wrong with Theory Dualism?*

In opposition to theory dualists, reductionists think we ought to strive for an integration of psychological and neurobiological theory. Obviously, a crucial element in the discussion concerns what is meant by "reduction"; hence, part of what must first be achieved is a proper account of what sort of business inter-theoretic reduction is.

Roughly, the account is this: Reductions are *explanations* of phenomena described by one theory in terms of the phenomena described by a more basic theory. Reductions typically involve the co-evolution of theories over time, and as they co-evolve, one theory is normally revised, corrected and modified by its co-evolutionary cohort theory at the other level. This revisionary interaction can, and usually does, go both ways; from the more basic to the less basic theory and vice versa. It is important to emphasize the modification to theories as they co-evolve, because sometimes the modification is radical and entails massive reconfiguration of the very categories used to describe the phenomena. In such an event, the very data to be explained may come to be redescribed under pressure from the evolving theories. Examples of categories that have undergone varying degrees of revision, from the minor to the radical, include impetus, caloric, gene, neuron, electricity, instinct, life, and very recently, excitability (in neurons) (Schaffner 1976, P. M. Churchland 1979, Hooker 1981).

Because reductionism is frequently misunderstood, it is necessary to be explicit about what is *not* meant. First, seeking reductions of macro-level theory to micro-level theory does not imply that one must first know everything about the elements of the micro theory before research at the macro-level can be usefully undertaken. Quite the reverse is advocated—research should proceed at all levels of the system, and co-evolution of theory may enhance progress at all levels. Data from one level *constrain* theorizing at that level and at other levels. Additionally, the reduction of theories does *not* mean that the reduced phenomena somehow disappear or are discredited. The theory of optics was reduced to the theory of electromagnetic radiation, but light itself did not disappear nor did it become disreputable to study light at the macro level. Nor was the reduced theory cast out as useless or discredited; on the contrary, it was and continues to be useful for addressing phenomena at a higher level of description. As for the phenomenon, it is what it is, and it continues to be whatever it is as theories are reduced or abandoned. Whether a category is ultimately rejected or revised depends on its scientific integrity, and that is, of course, determined empirically. (For more detail on inter-theoretic reduction, see P. S. Churchland 1986.)

Given this brief account of reduction as a backdrop, an outline of how the reductionist answers the theory dualist goes as follows:

Neuroscience *is* hard, but with many new techniques now available, an impressive body of data is available to constrain our theories, and a lot of data are very suggestive as to how neural networks function (see Sejnowski and Churchland, 1987). We have begun to see the shape of neurobiological answers to functional questions, such as how information is stored, how networks learn, and how networks of neurons represent.

*High-level states are multiply instantiable.* So what? If, in any given species, we can show that particular functional states are identical to specific neuronal configurations (for example, that being in REM sleep is having a specified neuronal state, or that one type of learning involves changing synaptic weights according to a Hebb rule), that will be sufficient to declare a reduction relative to that domain (Richardson 1979, Enc 1983, P. S. Churchland 1986; section 3 below). Very pure philosophers who cannot bring themselves to call these perfectly respectable domain-relative explanations "reductions" are really just digging in on who gets to use the word.

Moreover, it should be emphasized that the explanation of high-level cognitive phenomena will not be achieved directly in terms of phenomena at the lowest level of nervous-system organization, such as synapses and individual neurons. Rather, the explanation will refer to properties at higher structural levels, such as networks or systems. Functional properties of networks and systems will be explained by reference to properties at the next level down, and so on. What we envision is a chain of explanations linking higher to next-lower levels, and so on down the ladder of structural levels. (See Sejnowski and Churchland, in press.) The similarity of the information-processing function between two biological systems that are different at the level of the synaptic and cellular levels are probably a consequence of invariants that characterize dynamical systems in high-dimensional state spaces (Sejnowski, Koch, and Churchland 1988).

*Argument from intentionality.* A theory of how states in a nervous system represent or model the world will need to be set in the context of the evolution and development of nervous systems, and will try to explain the interactive role of neural states in the ongoing neuro-cognitive economy of the system. Nervous systems do not represent all aspects of the physical environment; they selectively represent information a species needs, given its environmental niche and its way of life. Nervous systems are programmed to respond to certain selected features, and within limits they learn other features through experience by encountering examples and generalizing. Cognitive neuroscience is now beginning to understand how this is done (Livingstone 1988; Goldman-Rakic 1988; Kelso, Ganong, and Brown 1986). Although the task is difficult, it now seems reasonable to assume that the "aboutness" or "meaningfulness" of representational states is not a spooky relation but a neurobiological relation. As we come to understand more about the dynamical properties of networks, we may ultimately be able to generate a theory of how human language is learned and represented by our sort of nervous system, and thence to explain language-dependent kinds of meaning.

Because this answer is highly cryptic and because intentionality has often seemed forever beyond the reach of neurobiology, the next section will focus on intentionality: the theory dualist's motivation, and the reductionist's strategy.

## 3. Levels, Intentionality, and the Sentence-Logic Model of the Mind

### *3.1 Sentential Attitudes and the Computer Metaphor*

Two deep and interrelated assumptions concerning the nature of cognition drive the third anti-reductionist argument:

Cognition essentially involves representations and computations. Representations are, in general, symbolic structures, and computations are, in general, rules (such as rules of logic) for manipulating those symbolic structures.

A good model for understanding mind-brain functions is the computer—that is, a machine based on the same logical foundations as a Turing machine and on the von Neumann architecture for a digital computer. Such machines are ideally suited for the manipulation of symbols according to rules. The computer metaphor suggests that the mind-brain, at the information processing level, can be understood as a kind of digital computer; the problem for cognitive psychology is to determine the program that our brains run.

The motivating vision here is that cognition is to be modeled largely on language and logical reasoning; having a thought is, functionally speaking, having a sentence in the head, and thinking is, functionally speaking, doing logic, or at least running on procedures very like logic. Put this baldly, it may seem faintly ridiculous, but the theory is supported quite plausibly by the observation that beliefs, thoughts, hopes, desires and so forth are essential in the explanation of cognition, and that such states are irreducibly semantic because they are identified in virtue of their content sentences. That is, such states are always and essentially beliefs that *p*, thoughts that *p*, or desires that *p*, where for *p* we substitute the appropriate sentence, such as "Nixon was a Russian spy" or "Custard is made with milk". Such cognitive states—the so-called sentential attitudes—are the states they are in virtue of the sentences that specify what they are about. Moreover, a content sentence stands in specific logical and semantic relations to other sentences. The state transitions are determined by semantic and logical relations between the content sentences, not by casual relations among states neurobiologically described. Thus, cognitive states have *meaning* (i.e. content, or intentionality), and it might be argued, that it is precisely in virtue of their meaningfulness that they play the role in cognition that they do.

The fundamental conception is, accordingly, well and truly rooted in folk psychology, the body of concepts and everyday lore by means of which we routinely explain one another's behavior by invoking sentential attitudes (Stich 1983, P. M. Churchland 1988)—e.g. Smith paid for the vase because he believed that his son had dropped it and he feared that the store owner would be angry. In these sorts of intentional explanations, the basic unit of representation is the sentence, and state transitions are accomplished through the following of rules: deductive inference, inductive inference, and assorted other rules.

Extending the framework of folk psychology to get an encompassing account of cognition in general, this approach takes it that thinking, problem solving, language use, perception, and so forth will be understood as we determine the sequence of sentences corresponding to the steps in a given information-processing task; i.e., as we understand the mechanics of sentence crunching. According to this research paradigm, known as sententialism, it is the task of cognitive science to figure out what programs the brain runs, and neuroscience can then check these top-down hypotheses against the wetware to see if they are generally possible. (See especially Fodor 1975, Fodor 1981, and Pylyshyn 1984.)

### *3.2. Is Cognition Mainly Symbol Manipulation in the Language of Thought?*

Although this view concerning the nature of cognition and the research strategy for studying cognition may be appealing (much of the appeal is derived from the comfortable place found for folk psychology), it suffers from major defects. Many of these defects have been discussed in detail by Anderson and Hinton (1981), P. S. Churchland (1986) and in various chapters of McClelland and Rumelhart (1986). A summary will call them to mind:

- Many cognitive tasks, such as visual recognition and answering simple true-or-false questions, can be accomplished in about half a second. Given what we know about conduction velocities and synaptic delays in neurons, this allows about 5 milliseconds per computational step, which means that there is time for only about 100 steps. For a sequential program run on a conventional computer, 100 steps is not going to get us remotely close to task completion. Feldman and Ballard (1982) call this the hundred-step rule.

- Anatomically and physiologically, the brain is a parallel system, not a sequential von Neumann machine. The neural architecture is highly interconnected. Neurons such as Purkinje cells may have upwards of 80,000 input connections, and neurons in cerebral cortex can have upwards of 10,000 output connections (Anderson and Hinton 1981, Pellionisz and Llinas 1982, Sejnowski 1986).
- However information is stored in nervous systems, it appears to be radically unlike information storage in a digital computer, where storage and processing are separated and items are stored in memory according to addressable *locations*. In nervous systems, information seems to be stored in the connections between the same neurons that process the information. There does not appear to be a distinct storage location for each piece of stored information, and information is content addressable rather than location addressable. Information storage is probably at least somewhat distributed rather than punctate, since memories tend to be degraded with damage to the system rather than selectively wiped out one by one.
- A task may fall gracefully on to one architecture, and not on to another. Certain kinds of tasks, such as numerical calculation, fall gracefully on to a von Neumann architecture, but others, such as learning or associative memory, do not. Things we humans find effortless (such as facial recognition and visual perception) are tasks which artificial intelligence has great difficulty simulating on a von Neumann architecture, whereas things we find it "effortful" to do (such as simple proofs in the propositional calculus or mathematical calculations) are straightforward for a digital computer (Anderson and Hinton 1981; Rumelhart, Hinton and McClelland 1986). This suggests that the computational style of nervous systems may be very unlike that suited to von Neumann architectures.
- The hardware-software analogy fails for many reasons, the most prominent of which are that nervous systems are plastic and that neurons continually change as we grow and learn. Related, perhaps, is the observation that nervous systems degrade gracefully and are relatively fault tolerant. A von Neumann machine is rigid and fault intolerant, and a breakdown of one tiny component disrupts the machine's performance.
- The analogy between levels of description in a conventional computer (such as the hardware-software distinction) and levels of explanation in nervous systems may well be profoundly misleading. Exactly how many of levels of organization we need to postulate

in order to understand nervous-system function is an empirical question, and it may turn out that there are many levels between the molecular and the behavioral. In nervous systems we may already discern as distinct descriptive levels the molecule, the membrane, the cell, the circuit, networks, the maps, brain systems, and several levels of behavior (from the reflexive to the highest levels of cognition). Other levels may come to be described as more is discovered about the nature of nervous systems. As is discussed below, the properties at one level may constrain the kind of properties realizable at another level.

- Nonverbal animals and infraverbal humans present a major problem for the sentence-logic theory of cognition: How is their cognition accomplished? On the sentence-logic theory of cognition, either their cognition resembles the human variety (and hence involves symbol manipulation according to rules and a language of thought replete with a substantial conceptual repertoire) or their cognitive processes are entirely different from the usual human ones. Neither alternative is remotely credible. The first lacks any evidence. At best, its defense is circular—it helps to save the theory. The second alternative entails a radical discontinuity in evolution—sufficiently radical that language-of-thought cognition is a bolt from the blue. This implies that evolutionary biology and developmental neurobiology are mistaken in some fundamental respects. Since neither alternative can be taken seriously, the hypothesis itself has diminished credibility.

If cognition, then, is *not*, in general, to be understood on the sentence-logic model, the pressing questions then are these: How *does* the brain represent? How do nervous systems model the external world of objects in motion and the internal world of the nervous system itself? And when representations do stand in semantic and logical relations to one another, how is this achieved by neural networks? How is the semantic and logical structure of language—as we both comprehend and speak—represented in the brain? According to the rejected model, we postulate an internal organization—a language of thought—*with the very same structure and organization as language*. But if that model is rejected, what do we replace it with?

These are, of course, *the* central questions, and getting answers will not be easy. But the difficulty should not make the language-of-thought hypothesis more appealing. In certain respects, the current

scientific state of a general theory of representation is analogous to the science of embryology in the nineteenth century. The development of highly structured, complex, fully-formed organisms from eggs and sperm is a profoundly amazing thing. Faced with this mystery, some scientists concluded that the only way to explain the emergence of a fully structured organism at birth was to assume that the structure was already there. Hence the homuncular theory of reproduction, which claimed that a miniature but complete human already exists in the sperm and merely expands during its tenure in the womb.

We now know that there *is* structure in the sperm (and the egg)—not in the form of a miniature, fully structured organism, but mainly in the form of DNA—a molecule that looks not at *all* like a fully formed human. Thus, the structure of the cause does not resemble the structure of the effect. Accordingly, the homuncular theorists were right in supposing that the highly structured neonate does not come from *nothing*, but they were wrong in looking for a structural resemblance between cause and effect. It was, of course, terribly hard to imagine the nature of the structural organization that enables development yet in no way resembles the final product. Only through molecular biology and detailed work in embryology have we begun to understand how one kind of structure can, through intermediate mechanisms, yield another, very different kind of structure.

The parallel with cognitive neurobiology is this: The neuronal processes underlying cognition have a structure of some kind, but almost certainly it will not, in general, look anything like the semantic/logic structure visible in overt language. The organizational principles of nervous systems are what permit highly complex, structured patterns of behavior, for it is certain that the behavioral structure does not emerge magically from neuronal chaos. As things stand, it is very hard to imagine what those organizational principles could look like, and, just as in genetics and embryology, we can find answers only by framing hypotheses and doing experiments.

Instead of starting from the old sentence-logic model, we model information processing in terms of *the trajectory of a complex nonlinear dynamical system in a very high-dimensional space.* This structure does not resemble sentences arrayed in logical sequences, but it is potentially rich enough and complex enough to yield behavior capable of supporting semantic and logical relationships. We shall now explore what representing looks like in a particular class of nonlinear dynamical systems called connectionist models.

## 4. Representation in Connectionist Models

As the name implies, a connectionist model is characterized by connections and differential strengths of connection between processing units. Processing units are meant to be like neurons and communicate with one another by signals (such as a firing rate) that are numerical rather than symbolic. Connectionist models are designed to perform a task by specifying the architecture: the number of units, their arrangement in layers and columns, the patterns of connectivity, and the weight or strength of each connection (figures 1 and 2). These models have close ties with the computational level on which the task is specified, and with the implementation level on which the task is physically instantiated (Marr 1982). This species of network models should properly be considered a class of algorithms specified at various levels of organization—in some cases at the small-circuit level, in other cases at the system level. Both the task description and the neural embodiment are, however, crucially important in constraining the class of networks that will be explored. On the one hand the networks have to be powerful enough to match human performance of the computational tasks, and on the other hand they have to be built from the available materials. In the case of the brain, that means neurons and synapses; in the case of network models, that means neuron-like processing units and synapse-like weights.

Digital computers are used to simulate neural networks, and the network models that can be simulated on current machines are tiny in comparison with the number of synapses and neurons in the mammalian brain. The networks that have been constructed should be understood, therefore, as small parts of a more complex processing system whose general configuration has not yet been worked out, rather than as simulations of a whole system. To avoid misunderstanding, it should be emphasized that connectionist models cannot yet support a full cognitive system. To begin to reach that goal will require both a computing technology capable of supporting more detailed simulations and a more complete specification of the nervous system.

Granting these limitations, we may nonetheless be able to catch a glimpse of what representations might look like within the parallel-style architecture of the brain by taking a look inside a connectionist network. The place to look is in the dynamics of the system; that

is, in the patterns of activity generated by the system of interconnected units. This approach has its roots in the work of previous generations of researchers–primarily the gestalt school of psychology and D. O. Hebb (1949), who developed many ideas about learning and representation in neural assemblies. Only recently, however, has sufficient computer power been available to explore the consequences of these ideas by direct simulation, since the dynamics of massively parallel nonlinear networks is highly computation intensive. Parallel-network models are now being used to explore many different aspects of perception and cognition, (McClelland and Rumelhart 1986; Feldman and Ballard 1982; *Cognitive Science*, volume 9, special issue), but in this chapter we shall focus on two representative examples. The first is NETtalk, one of the most complex network models yet constructed, which learns to convert English text to speech sounds (Sejnowski and Rosenberg 1987, 1988). The second is a network model that computes surface curvatures of an object from its gray-level input image. NETtalk will be used primarily to illustrate two things: how a network can learn to perform a very complex task without symbols and without rules to manipulate symbols, and the differences between local and distributed representations.

Connectionist models can be applied on a large scale to model whole brain systems or, on a smaller scale, to model particular brain circuits. NETtalk, is on a large scale, since the problem of pronunciation is constrained mainly by the abstract cognitive considerations and since its solution in the brain must involve a number of systems, including the visual system, the motor-articulatory system, and the language areas. The second example is more directly related to smaller brain circuits used in visual processing; the representational organization achieved by the network model can be related to the known representational organization in visual cortex.

In the models reviewed here, the processing units sum the inputs from connections with other processing units, each input weighted by the strength of the connection. The output of each processing unit is a real number that is a nonlinear function of the linearly summed inputs. The output is small when the inputs are below threshold, and it increases rapidly as the total input becomes more positive. Roughly, the activity level can be considered the sum of the postsynaptic potentials in a neuron, and the output can be considered its firing rate (figure 1).

*4.1. Speech Processing: Text to Speech.*

In the simplest NETtalk system[2] there are three layers of processing units. The first level receives as input letters in a word; the final layer yields the elementary speech sounds, or phonemes (table 1); and an intervening layer of "hidden units" which is fully connected with the input and output layers, performs the transformation of letters to sounds (figure 3). On the input layer, there is *local representation* with respect to letters because single units are used to represent single letters of the alphabet. Notice, however, that the representation could be construed as *distributed* with respect to *words,* inasmuch as each word is represented as a pattern of activity among the input units. Similarly, each phoneme is represented by a pattern of activity among the output units, and phonemic representation is therefore distributed. But each output unit is coded for a particular *distinctive feature* of the speech sound, such as whether the phoneme was voiced, and consequently each unit is local with respect to distinctive features.

NETtalk has 309 processing units and 18,629 connection strengths (weights) that must be specified. The network does not have any initial or built-in organization for processing the input or (more exactly) mapping letters onto sounds. All the structure emerges during the training period. The values of the weights are determined by using the "back-propagation" learning algorithm developed by Rumelhart, Hinton, and Williams (1986). (For reviews of network learning algorithms, see Hinton, 1988, and Sejnowski, 1988.) The strategy exploits the calculated error between the *actual* values of the processing units in the output layer and the *desired* values, which is provided by a training signal. The resulting error signal is propagated from the output layer backward to the input layer and used to adjust each weight in the network. The network learns, as the weights are changed, to minimize the mean squared error over the training set of words. Thus, the system can be characterized as following a path in weight space (the space of all possible weights) until it finds a minimum (figure 4). The important point to be illustrated, therefore, is this: The network processes information by nonlinear dynamics, not by manipulating symbols and accessing rules. It learns by gradient descent in a complex interactive system, not by generating new rules (Hinton and Sejnowski 1986).

The issue that we want to focus on next is the structural

organization that is "discovered" by the network, in virtue of which it succeeds in converting letters to phonemes and manages to pronounce, with few errors, the many irregularities of English. If there are no rules in the network, how is the transformation accomplished? Since a trained network can generalize quite well to new words, some knowledge about the pattern of English pronunciation must be contained inside the network. Although a representational organization was imposed on the input and output layers, the network had to create new, internal representations in the hidden layer of processing units. How did the network organize its "knowledge"? To be more accurate: How did the equivalence class of networks organize its knowledge? (Each time the network was started from a random set of weights, a different network was generated.)

The answers were not immediately available, because a network does not leave an explanation of its travels through weight space, nor does it provide a decoding scheme when it reaches a resting place. Even so, some progress was made by measuring the activity pattern among the hidden units for specific inputs. In a sense, this test mimics at the modeling level what neurophysiologists do at the cellular level when they record the activity of a single neuron to try to find the effective stimulus that makes it respond. NETtalk is a fortunate "preparation", inasmuch as the number of processing units is relatively small, and it is possible to determine the activity patterns of all the units for all possible input patterns. These measurements, despite the relatively small network, did create a staggering amount of data, and then the puzzle was this: How does one find the order in all this data?

For each set of input letters, there is a pattern of activity among the hidden units (figure 5). The first step in the analysis of the activity of the hidden units was to compute the average level of activity for each letter-to-sound correspondence. For example, all words with the letter c in the middle position yielding the hard-c sound /k/ were presented to the network, and the average level of activity was calculated. Typically, about 15 of the 80 hidden units were very highly activated on average, and the rest of the hidden units had little or no activity. This procedure was repeated for each of the 79 letter-to-sound correspondences. The result was 79 vectors, each vector pointing in a different direction in the 80-dimensional space of average hidden-unit activities. The next step was to explore the relationship among the vectors in this space by cluster analysis. It

is useful to conceive of each vector as the internal code that is used to represent a specific letter-to-sound correspondence; consequently, those vectors that clustered close together would have similar codes.

Remarkably, all the vectors for vowel sounds clustered together, indicating that they were represented in the network by patterns of activity in units that were distinct from those representing the consonants (which were themselves clustered together). (See figure 6.) Within the vowels, all the letter-to-sound correspondences that used the letter a were clustered together, as were the vectors of e, i, o, and u and the relevant instances of y. This was a very robust organizational scheme that occurred in all the networks that were analyzed, differences in starting weights notwithstanding. The coding scheme for consonants was more variable from network to network, but as a general rule the clustering was based more on similarities in sounds than on letters. Thus, the labial stops /p/ and /b/ were very close together in the space of hidden-unit activities, as were all the letter-to-sound correspondences that result in the hard-c sound /k/.

Other statistical techniques, such as factor analysis and multi-dimensional scaling, are also being applied to the network, and activity patterns from individual inputs, rather than averages over classes, are also being studied (Rosenberg 1988). These statistical techniques are providing us with a detailed description of the representation for single inputs as well as classes or input-output pairs.

Several aspects of NETtalk's organization should be emphasized:

- The representational organization visible in the trained-up network is not programmed or coded into the network; it is found by the network. In a sense it "programs" itself, by virtue of being connected in the manner described and having weights changed by experience according to the learning algorithm. The dynamical properties of this sort of system are such that the network will settle into the displayed organization.
- The network's representation for letter-to-sound correspondences is neither local nor completely distributed; it is somewhere in between. The point is, each unit participates in more than one correspondence, and so the representation is not local, but since it does not participate in all correspondences, the representation is not completely distributed either.
- The representation is a property of the collection of hidden units, and does not resemble sentence-logic organization.

- The organization is structured, which suggests that emergent subordinate and superordinate relations might be a general principle of network organization that could be used as input for other networks assigned other tasks, if NETtalk were embedded in a larger system of networks.
- General properties of the hierarchical organization of letter-to-sound correspondences emerged only at the level of groups of units. This organization was invariant across all the networks created from the same sample of English words, even where the processing units in distinct networks had specialized for a different aspect of the problem.
- Different networks created by starting from different initial conditions all achieved about the same level of performance, but the detailed response properties of the individual units in the networks differed greatly. Nonetheless, all the networks had similar functional clusterings for letter-to-sound correspondences (figure 6). This suggests that single neurons only code information relative to other neurons in small groups or assemblies (Hebb 1949).

The representational organization in NETtalk may illustrate important principles concerning network computation and representation, but what do they tell us about neural representations? Some of the principles uncovered might be generally applicable to a wide class of tasks, but it would be surprising if the details of the model bore any significant resemblance to the way reading skills are represented in the human nervous system. NETtalk is more of a demonstration of certain network capacities and properties than a faithful model of some subsystem of the brain, and it may be a long time before data concerning the human neurobiology of reading become available. Nevertheless, the same network techniques that were used to explore the language domain can be applied to problems in other domains, such as vision, where much more is known about the anatomy and the physiology.

### *4.2. Visual Processing: Computing Surface Curvature from Shaded Images*

The general constraints from brain architecture touched on in section 3 should be supplemented, wherever possible, by more detailed constraints from brain physiology and anatomy. Building models of real neural networks is a difficult task, however, because

essential knowledge about the style of computation in the brain is not yet available (Sejnowski 1986). Not only is the fine detail (such as the connectivity patterns in neurons in cerebral cortex) not known, but even global-level knowledge specifying the flow of information through different parts of the brain during normal function is limited. Even if more neurophysiological and neuroanatomical detail were available, current computing technology would put rather severe limits on how much detail could be captured in a simulation. Nevertheless, the same type of network model used in NETtalk could be useful in understanding how information is coded within small networks confined to cortical columns. The processing units in this model will be identified with neurons in the visual cortex.

Ever since Hubel and Wiesel (1962) first reported that single neurons in the cat visual cortex respond better to oriented bars of lights and to dark/light edges than to spots of light, it has been generally assumed, or at least widely hoped, that the function of these neurons is to detect boundaries of objects in the world. In general, the inference from a cell's response profile to its function in the wider information-processing economy is intuitively very plausible, and if we are to have any hope of understanding neural representations we need to start in an area—such as visual cortex—where it is possible to build on an impressive body of existing data. The trouble is, however, that many functions are consistent with the particular response properties of a neuronal population. That a cell responds optimally to an oriented bar of light is compatible with its having lots of functions other than detecting object boundaries, though the hypothesis that it serves to detect boundaries does tend to remain intuitively compelling. To see that our intuitions might really mislead us as we try to infer function from response profiles, it would be useful if we could demonstrate this point concretely. In what follows we shall show how the same response properties could in fact serve in the processing of visual information about the regions of a surface between boundaries rather than about the boundaries themselves.

Boundaries of objects are relatively rare in images, yet the preponderance of cells in visual cortex respond preferentially to oriented bars and slits. If we assume that all those cells are detecting boundaries, then it is puzzling that there should be so many cells whose sole function is to detect boundaries when there are not many boundaries to detect. It would, therefore, seem wasteful if, of all the neurons with oriented fields, only a small fraction carried useful

information about a particular image. Within their boundaries, most objects have shaded or textured surfaces that will partially activate these neurons. The problem accordingly is this: Can the information contained in a population of partially-activated cortical neurons be used to compute useful information about the three-dimensional surfaces between the boundaries of objects in the image?

One of the primary properties of a surface is its curvature. Some surfaces, such as the top of a table, are flat, and have no intrinsic curvature. Other surfaces, such as cylinders and spheres, are curved, and around each point on a surface the degree of curvature can be characterized by the direction along the surface of maximum and minimum curvature. It can be shown that these directions are always at right angles to each other, and the values are called the principle curvatures (Hilbert and Cohn-Vossen 1952). The principal curvatures and the orientation of the axes provide a complete description of the local curvature.

One problem with extracting the principal curvatures from an image is that the gray-level shading depends on many factors, such as the direction of illumination, the reflectance of the surface, and the orientation of the surface relative to the viewer. Somehow our visual system is able to separate these variables and to extract information about the shape of an object independent of these other variables. Pentland (1984) has shown that a significant amount of information about the curvature of a surface is available locally. Can a network model be constructed that can extract this information from shaded images?

Until recently it was not obvious how to begin to construct such a network, but network learning algorithms (see above) provide us with a powerful method for creating a network by giving it examples of the task at hand. The learning algorithm is being used in this instance simply as a design tool to see whether some network can be found that performs the task. Many examples of simple surfaces (elliptic paraboloids) were generated and presented to the network. A set of weights was indeed found with this procedure that, independent of the direction of illumination, extracted the principal curvatures of three-dimensional surfaces and the direction of maximum curvature from shaded images (Lehky and Sejnowski 1988a,b).

The input to the network is from an array of on-center and off-center receptive fields similar to those of cells in the lateral geniculate

nucleus. The output layer is a population of units that conjointly represent the curvatures and the broadly tuned direction of maximum curvature. The units of the intermediate layer, which are needed to perform the transformation, have oriented receptive fields, similar to those of simple cells in the visual cortex of cats and monkeys that respond optimally to oriented bars and edges (figure 7). It is important to emphasize that these properties of the hidden units were not put into the network directly but emerged during training. The system "chose" these properties because they are useful in performing a particular task. Interestingly, the output units, which were required to code information about the principal curvatures and principal orientations of surfaces, had properties, when probed with bars of light, that were similar to those of a class of complex cells that are end-stopped (Lehky and Sejnowski 1988a,b). The surprising thing, given the plausible receptive-field-to-function inference rule, is that the function of the units in the network is not to detect bounding contours, but to extract curvature information from shaded images. Whether or not curvature is directly represented in visual cortex can be tested by designing experiments with images of curved surfaces.

What the shape-from-shading network demonstrates is that we cannot directly infer function from receptive field properties. In the trained-up network the hidden units represent an intermediate transformation for a computational task quite different from the one that has been customarily ascribed to simple cells in visual cortex—they are used to determine shape from shading, not to detect boundaries. It turns out, however, that the hidden units have *receptive fields similar to those of simple cells in visual cortex.* Therefore, bars and edges as receptive-field properties do not necessarily mean that the cell's function is to detect bars and edges in objects; it might be to detect curvature and shape, as it is in the network model, or perhaps some other surface property such as texture. The general implication is that there is no way of determining the function of each hidden unit in the network simply by "recording" the receptive-field properties of the unit. This, in turn, implies that, despite its intuitive plausibility, the receptive-field-to-function inference rule is untenable.

The function of a unit is revealed only when its *outputs*—its "projective field", (Lehky and Sejnowski, 1988a,b)—are also examined. It is the projective field of a unit that provides the information needed to interpret the unit's computational role in the network. In

the network model the projective field could be examined directly, but in real neural networks, it can only be inferred indirectly by examining the next stage of processing.

### *4.3. Next-Generation Networks*

NETtalk and the shape-from-shading network are important examples because they yield clues to how the nervous system can embody models of various domains of the world. Parallel-network modeling is still in a pioneering stage of development. There are bound to be many snags and hitches, and many problems yet undreamt of will have to be solved. At this stage, the representational structure of networks has not yet been explored in detail, nor is it known how well the performance of network models will scale with the number of neurons and the difficulty of the task. (That is, will representations and computations in a cortical column with 200,000 neurons be similar to those in a model network comprising only a few hundred processing units?)

Moreover, taken literally as a model of functioning neurons, back-propagation is biologically implausible, inasmuch as error signals cannot literally be propagated back down the very same axon the signal came up. Taken as a *systems-level* algorithm, however, back-propagation may have a realization using feedback projections that do map onto neural hardware. Even squarely facing these cautionary considerations, the important thing is that something with this sort of character at least lets us see what representational structure—good, meaty, usable structure—could *look like* in a neuronal network.

Temporal chaining of sequences of representations is probably a prominent feature of many kinds of behavior, and it may turn out to be particularly important for language acquisition and use. It is conceivable that structured sequences—long, temporally extended sequences—are the elements of an abstract sort of neural state space that enable humans to use language. Sereno (1986) has suggested something along these lines, pointing out that DNA, as a spatially extended sequence of nucleotides, allows for encoding; by analogy, one may envision that the development of mechanisms for generating temporally extended sequences of neuronal (abstract) structures may allow for a kind of structured behavior (i.e. language) that short sequences do not allow for. (See also MacKay 1987; Dehaene et al. 1987.)

One promising strategy will be to try first to unscramble the more fundamental kinds of representing accomplished by nervous systems, shelving until later the problem of complex representations such as linguistic representations. To solve such problems, the solutions discovered for simpler representations may be crucial. At the most basic level, there appears to be an isomorphism between cell responses and external events (for example cells in visual cortex responding to bars of light moving in a specific direction). At higher levels the receptive field properties change (Allman et al. 1985, Andersen 1987), and it may be that the lower-level isomorphism gives way to more complicated and dynamic network effects. Motivation, planning, and other factors may at this level, have roles in how a representation is generated. At still higher levels, still other principles may be operative. Once we understand the nature of representing in early sensory processing, as we have indeed begun to do, and go on to address the nature of representations at more and more abstract levels, we may finally be able to address how learning a language yields another kind of representation, and how symbols can be represented in neural networks. Whatever the basic principles of language representation, they are not likely to be utterly unrelated to the way or ways that the nervous system generates visual representations or auditory representations, or represents spatial maps and motor planning. (On semantic relations in connectionist models, see Hinton 1981, 1986.)

## 5. Dogmas and Dreams: George Boole, Ramon y Cajal, David Marr

The connectionist models discussed are valuable for the glimpse of representational and computational space that they provide, for it is exactly such glimpses that free us from the bonds of the intuitive conceptions of representation as language-like and computation as logic-like. They thus free us from what Hofstadter (1982) called the *Boolean Dream*, where all cognition is symbol-manipulation according to the rules of logic.

Equally important, they also free us from what we call the *Neurobiologists' Dream* (perhaps, with all due respect, it might be called Cajal's Dream), which is really the faith that the answers we seek will be manifest once the fine-grain details of each neuron (its

morphology, physiology, and connections) are revealed. These models also teach the tremendously important lesson that *system properties are not accessible at the single unit level.* In a system, what we need to know is how the elements in large set of elements interact over time. Until we have new physiological techniques for supplying data of that sort, building network models is a method of first resort.

To be really useful, a model must be biologically constrained. However, exactly which biological properties are crucial to a model's utility and which can be safely ignored until later, are matters that can be decided only by hunches until a mature theory is in place. Such 'bottom-up' constraints are crucial, since computational space is immensely vast, too vast for us to be lucky enough to light on the correct theory simply from the engineering bench. Moreover, the brain's solutions to the problems of vision, motor control, and so forth may be far more powerful, more beautiful, and even more simple than what we engineer into existence. This is the point of Orgel's Second Rule: Nature is more ingenious than we are. And we stand to miss all that power and ingenuity unless we attend to neurobiological plausibility. The point is, *evolution has already done it*, so why not learn how that stupendous machine, our brain, actually works?

This observation allows us to awake from *Marr's Dream* of three levels of explanation: the computational level of abstract problem analysis, the level of the algorithm, and the level of physical implementation of the computation. In Marr's view, a higher level was independent of the levels below it, and hence computational problems could be analyzed independently of an understanding of the algorithm that executes the computation, and the algorithmic problem could be solved independently of an understanding of the physical implementation. Marr's assessment of the relations between levels has been reevaluated, and the dependence of higher levels on lower levels has come to be recognized.

The matter of the interdependence of levels marks a major conceptual difference between Marr and the current generation of connectionists. Network models are not independent of either the computational level or the implementational level; they depend in important ways on constraints from all levels of analysis. Network models show how knowledge of brain architecture can contribute to the devising of likely and powerful algorithms that can be efficiently implemented in the architecture of the nervous system and may alter even how we construe the computational problems.

On the heels of the insight that the use of constraints from higher up and lower down matters tremendously, the notion that there are basically *three* levels of analysis also begins to look questionable. If we examine more closely how the three levels of analysis are meant to map onto the organization of the nervous system the answer is far from straightforward.

To begin with, the idea that there is essentially one single implementational level is an oversimplification. Depending on the fineness of grain, research techniques reveal structural organization at many strata: the biochemical level; then the levels of the membrane, the single cell, and the circuit; and perhaps yet other levels such as brain subsystems, brain systems, brain maps, and the whole central nervous system. But notice that at each structurally specified stratum we can raise the functional question: What does it contribute to the wider, functional business of the brain?

This range of structural organization implies, therefore, that the oversimplification with respect to implementation has a companion over-simplification with respect to computational descriptions. And indeed, on reflection it does seem most unlikely that a single type of computational description can do justice to the computational niche of diverse structural organization. On the contrary, one would expect distinct task descriptions corresponding to distinct structural levels. But if there is a ramifying of task specifications to match the ramified structural organization, this diversity will probably be reflected in the ramification of the *algorithms* that characterize how a task is accomplished. And this, in turn, means that the notion of *the* algorithmic levels is as oversimplified as the notion of *the* implementation level.

Similar algorithms were used to specify the network models in NETtalk and the shape-from-shading network, but they have a quite different status in these two examples. On this perspective of the levels of organization, NETtalk is a network relevant to the *systems* level, whereas the shape-from-shading network is relevant to the *circuit* level. Since the networks are meant to reflect principles at entirely different levels of organization, their implementations will also be at different scales in the nervous system. Other computational principles may be found to apply to the single cell or to neural maps.

Once we look at them closely, Marr's three *levels of analysis* and the brain's *levels of organization* do not appear to mesh in a very useful or satisfying manner. So poor is the fit that it may be doubted

whether levels of analysis, *as conceived by Marr,* have much methodological significance. Accordingly, in light of the flaws with the notion of *independence*, and in light of the flaws with the *tripartite* character of the conception levels, it seems that Marr's dream, inspiring though it was for a time, must be left behind.

The vision that inspires network modeling is essentially and inescapably interdisciplinary. Unless we explicitly theorize above the level of the single cell, we will never find the key to the order and the systematicity hidden in the blinding minutiae of the neuropil. Unless our theorizing is geared to mesh with the neurobiological data, we risk wasting our time exploring some impossibly remote, if temporarily fashionable, corner of computational space. Additionally, without the constraints from psychology, ethology and linguistics to specify more exactly the parameters of the large-scale capacities of nervous systems, our conception of the functions for which we need explanation will be so woolly and tangled as to effectively smother progress.

Consequently, cross-disciplinary research, combining constraints from psychology, neurology, neurophysiology, linguistics, and computer modeling, is the best hope for the co-evolution that could ultimately yield a unified, integrated science of the mind-brain. It has to be admitted, however, that this vision is itself a dream. From within the dream, we cannot yet reliably discern what are the flaws that will impede progress, what crucial elements are missing, or at which points the vague if tantalizing hunches might be replaced by palpable results.

## Notes

1. An earlier exploration of these ideas is to be found in Kenneth Craik's book *The Nature of Explanation* (Cambridge University Press, 1943).
2. NETtalk networks can differ in how input letters and output phonemes are represented, and in the number and arrangement of hidden units.

## References

Allman, J., F. Miezin, and E. McGuinness. 1985. "Stimulus specific responses from beyond the classic receptive field." *Annual Review of Neuroscience* 8: 407-430.

Anderson, J. A., and G. E. Hinton. 1981. "Models of information processing in the brain." In Hinton and Anderson 1981.

Andersen, R. A. 1987. "The role of posterior parietal cortex in spatial perception and visual-motor integration." *Handbook of Physiology—The Nervous System V.* ed. V. B. Mountcastle, F. Plum and S. R. Geiger.

Chisholm, R. M. 1966. *Theory of Knowledge.* Englewood Cliffs, N.J.: Prentice-Hall.

Churchland, P. M. 1979. *Scientific Realism and the Plasticity of Mind.* Cambridge University Press.

Churchland, P. M. 1988. *Matter and Consciousness.* Cambridge, Mass.: MIT Press.

Churchland, P. S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* Cambridge, Mass.: MIT Press.

Davidson, D. 1974. "On the very idea of a conceptual scheme." *Proceedings and Addresses of the American Philosophical Association* 47: 5-20.

Dehaene, S., J.-P. Changeux, and J.-P. Nadal. 1987. "Neural networks that learn temporal sequences by selection", Proceedings of the National Academy of Sciences USA 84: 2727-2731.

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, Mass.: MIT Press.

Eccles, J. C. 1977. Part II of K. Popper, *The Self and Its Brain* (Berlin: Springer-Verlag). Enc, B. 1983. "In defense of the identity theory." *Journal of Philosophy* 80: 279-298.

Feldman, J. A. and F. H. Ballard. 1982. "Connectionist models and their properties." Cognitive Science 6: 205-254.

Fodor, J. A. 1975. *The Language of Thought.* New York: Crowell. (Paperback edition: Cambridge, Mass.: MIT Press, 1979)

Fodor, J. A. (1981). *Representations.* Cambridge, Mass.: MIT Press.

Goldman-Rakic, P. S. 1987. "Circuitry of primate prefrontal cortex and regulation of behavior by representational memory." In *Handbook of Physiology—The Nervous System V,* ed. V. B. Mountcastle, F. Plum, and S. R. Geiger.

Hebb, D. O. (1949). *Organization of Behavior.* New York: Wiley.

Hilbert, J. and S. Cohn-Vossen. 1952. *Geometry and the Imagination.* New York: Chelsea.

Hinton, G. E. 1981. "Implementing semantic networks in parallel hardware." In Hinton and Anderson 1981.

Hinton, G. E. 1986. "Learning distributed representations of concepts." In *Proceedings of the Eight Annual Conference of the Cognitive Science Society.* Hillsdale, N.J.: Erlbaum.

Hinton, G. E. (1988). "Connectionist learning procedures.", *Artificial Intelligence* (in press).

Hinton, G. E. and J. A. Anderson, eds. (1981). *Parallel models of associative memory.* Hillsdale, N.J.: Erlbaum.

Hinton, G. E. and T. J. Sejnowski (1986). "Learning and relearning in Boltzmann machines.: In: McClelland and Rumelhart.

Hofstadter, D. R. (1982). "Artificial Intelligence: Subcognition as computation." *Technical Report No. 132* Computer Science Department, Indiana University.

Hooker, C. A. (1981). "Toward a general theory of reduction. Part I: Historical and scientific setting . Part II: Identity in reduction. Part III: Cross-categorical reduction." *Dialogue* 20:38-59, 201-236, 496-529.

Hubel, D. H. and T. N. Wiesel (1962). "Receptive fields, binocular interaction and functional architecture in cat's visual cortex." *Journal of Physiology.* 160: 106-154.

Kelso, S. R., A. H. Ganong, and T. H. Brown. 1986. "Hebbian synapses in hippocampus". Proceedings of the National Academy of Sciences 83: 5326-5330.

Lehky, S. and T. J. Sejnowski (1988a). "Network model of shape-from-shading: Neural function arises from both receptive and projective fields" *Nature* 333:452-454.

Lehky, S. and T. J. Sejnowski (1988b). Neural network model for the representation of surface curvature from images of shaded surfaces." In: *Organizing Principles of Sensory Processing* ed. J. Lund (Oxford University Press).

Livingstone, M. S. 1988. "Art, illusion, and the visual system." *Scientific American* 258: 78-85.

MacKay, D. (1987). *The Organization and Perception of Action*, Berlin: Springer-Verlag.

Mandt, A. J. (1986). "The triumph of philosophical pluralism? Notes on the transformation of academic philosophy." *Proceedings and Addresses of the American Philosophical Association* 60:265-277.

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

McClelland, J. L. and D. E. Rumelhart (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, Mass.: MIT Press.

McGinn, C. (1982). *The Character of Mind.* Oxford: Oxford University Press.

Pellionisz, A. and R. Llinas (1982). "Space-time representation in the brain. The cerebellum as a predictive space-time metric tensor." *Neuroscience* 7: 2249-2970.

Pentland, A. P. (1984). "Local shading analysis". *IEEE Transactions: Pattern Analysis and Machine Intelligence* 6: 170-187.

Putnam, H. (1967). "The nature of mental states." *In: Arts, mind and religion,* ed. W. H. Capitan and D. D. Merrill, 37-48. Pittsburgh: University of Pittsburgh Press. Reprinted in *Mind, Language and Reality: Philosophical Papers*, Vol. 2, Hilary Putnam (1975). Cambridge: Cambridge University Press.

Pylyshyn, Z. (1984). *Computation and cognition.* Cambridge, Mass.: MIT Press.

Quine, W. V. O. (1960). *Word and object.* Cambridge, Mass.: MIT Press.

Richardson, R. (1979). "Functionalism and reductionism". Philosophy of Science 46: 533-558.

Rosenberg, C. R. (1988). Ph.D. Thesis, Princeton University.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. (1986). "Learning internal representations by error propagation.: In: McClelland and Rumelhart.

Schaffner, K. F. (1976). "Reductionism in biology: Prospects and problems." In *PSA Proceedings* 1974, ed. R. S. Cohen, C. A. Hooker, A. C. Michalos, and J. W. Van Evra. Dordrecht: Reidel.

Sejnowski, T. J. (1986) "Open questions about computation in cerebral cortex." In McClelland and Rumelhart.

Sejnowski, T. J. (1988). "Neural network learning algorithms", In: *Neural Computers*, R. Eckmiller and C. von der Malsberg (eds.), Berlin: Springer Verlag.

Sejnowski, T. J. and Rosenberg, C. R. (1987) "Parallel networks that learn to pronounce English text." *Complex Systems.* 1: 145-168.

Sejnowski, T. J. and Rosenberg, C. R. (1988). "Learning and Representation in Connectionist models." In: *Perspective in memory research and training* ed. M. Gazzaniga. Cambridge, Mass.: MIT Press.

Sejnowski, T. J., and P. S. Churchland. (In press). "Brain and Cognition." In: *Foundations of Cognitive Science*, ed. M. I. Posner (Cambridge, Mass.: MIT Press).

Sejnowski, T. J., C. Koch, and P. S. Churchland. 1988. "Computational neuroscience." *Science* 241, 1299-1306.

Sereno, M. (1986). "A program for the neurobiology of mind." *Inquiry* 29: 217-240.

Skinner, B. F. (1957). *Verbal behavior.* New York: Appelton-Century-Crofts.

Skinner, B. F. (1976). *About behaviorism.* New York: Knopf.

Smart, J. J. C. (1959). "Sensations and brain processes." *Philosophical Review* 68:141-56.

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, Mass.: MIT Press.

Strawson, P. F. (1966). *The bounds of sense: An essay on Kant's Critique of Pure Reason.* London: Methuen.

Swinburne, R. (1986) *The evolution of the soul.* Oxford: Oxford University Press.

**Table and Figure Legends**

Table 1: List of phonemes used in NETtalk and examples (italicized letters).

Fig. 1: Left: Schematic model of a neuron-like processing unit that receives synapse-like inputs from other processing units. Right: Nonlinear sigmoid-shaped transformation between summed inputs and the output "firing rate" of a processing unit. The output is a continuous value between 0 and 1.

Fig. 2: Schematic model of a three-layered network. Each input unit makes connections with each of the hidden units on the middle layer, which in turn projects to each of the output units. This is a feedforward architecture in which information provided as an input vector flows through the network, one layer at a time, to produce an output vector. More complex architectures allow feedback connections from an upper to a lower layer and lateral interactions between units within a layer.

Fig. 3: Schematic drawing of the NETtalk network architecture. A window of letters in an English text is fed to an array of 203 input units

arranged in 7 groups of 29 units each. Information from these units is transformed by an intermediate layer of 80 hidden units to produce a pattern of activity in 26 output units. The connections in the network are specified by a total of 18,629 weight parameters (including a variable threshold for each unit). During the training, information about the desired output provided by the Teacher is compared with the actual output of the network, and the weights in the network are changed slightly so as to reduce the error.

Fig. 4: Schematic drawing of a path followed in weight space as the network finds a minimum of the average error over the set of training patterns. Only two weights out of many thousands are shown. The learning algorithm only ensures convergence to a local minimum, which is often a good solution. Typically, many sets of weights are good solutions, so the network is likely to find one of them from a random starting position in weight space. The learning time can be reduced by starting the network near a good solution; for example, the pattern of connections can be limited to a geometry that reduces the number of variable weights that must be searched by gradient descent.

Fig. 5: Levels of activation in the layer of hidden units for a variety of words. The input string in the window of seven letters is shown to the left, with the target letter emphasized. The output from the network is the phoneme that corresponds to the target letter. The transformation is accomplished by 80 hidden units, whose activity levels are shown to the right in two rows of 40 units each. The area of each white square is proportional to the activity level. Most units have little or no activity for a given input, but a few are highly activated.

Fig. 6: Hierarchical clustering of hidden units for letter-to-sound correspondences. The vectors of average hidden unit activity for each correspondence ("l"-p for letter "l" and phoneme p) were successively merging from right to left in the binary tree. The scale at the top indicates the Euclidean instance between the clusters. (From Sejnowski and Rosenberg 1987.)

Fig. 7: Hinton diagram showing the connection strengths in a network that computes the principal curvatures and direction of minimum curvature from shaded images in a small patch of the visual field corresponding roughly to the area represented in a cortical column. There are 12 hidden units which receive connections from the 122 inputs and project to each of the 23 output units. The diagram shows each of the connection strengths to and from the hidden units. Each weight is represented by one square, the area of which is proportional to the magnitude of the weight. The color is white if the weight is excitatory and black if it is inhibitory. The inputs are two hexagonal arrays of 61 processing units each. Each input unit has a concentric on-center (top) or off-surround (bottom) receptive field similar to those of principal cells in the lateral geniculate nucleus. The output consists of 24 units that conjointly

represent the direction of maximum curvature (six columns) and principal curvature (four rows: two for each principal curvature.) Each of the 12 hidden units is represented in the diagram in a way that reveals all the connections to and from the unit. Within each of the 12 gray background regions, the weights from the inputs are shown on the bottom and the weights to the output layer are shown above. To the left of each hidden unit, the lone square gives the threshold of the unit, which was also allowed to vary. Note that there emerged two different types of hidden units as revealed by the "projective field". The six units in the bottom row and the fourth and fifth from the left in the top row were mainly responsible for providing information about the direction of minimum curvature, while others were responsible for computing the signs and magnitudes of the two principal curvatures. The curvature-selective units could be further classified as convexity detectors (top row, third from left) or elongation filters (top row, second and sixth from left).

Table 1

*Symbols for Phonemes*

| Symbol | Phoneme | Symbol | Phoneme |
|---|---|---|---|
| /a/ | f*a*ther | /C/ | *ch*in |
| /b/ | *b*et | /D/ | *th*is |
| /c/ | b*ough*t | /E/ | bet |
| /d/ | *d*ebt | /G/ | si*ng* |
| /e/ | bake | /I/ | b*i*t |
| /f/ | *f*in | /J/ | *g*in |
| /g/ | *g*uess | /K/ | sexual |
| /h/ | *h*ead | /L/ | bott*le* |
| /i/ | Pete | /M/ | absy*m* |
| /k/ | *K*en | /N/ | butto*n* |
| /l/ | *l*et | /O/ | b*oy* |
| /m/ | *m*et | /Q/ | *qu*est |
| /n/ | *n*et | /R/ | b*ir*d |
| /o/ | b*oa*t | /S/ | *sh*in |
| /p/ | *p*et | /T/ | *th*in |
| /r/ | *r*ed | /U/ | b*oo*k |
| /s/ | *s*it | /W/ | b*ou*t |
| /t/ | *t*est | /X/ | excess |
| /u/ | l*u*te | /Y/ | c*u*te |
| /v/ | *v*est | /Z/ | leisure |
| /w/ | *w*et | // | b*a*t |
| /x/ | *a*bout | /!/ | Nazi |
| /y/ | *y*et | /#/ | examine |
| /z/ | *z*oo | /*/ | *o*ne |
| /A/ | b*i*te | /|/ | log*i*c |
| | | /^/ | b*u*t |

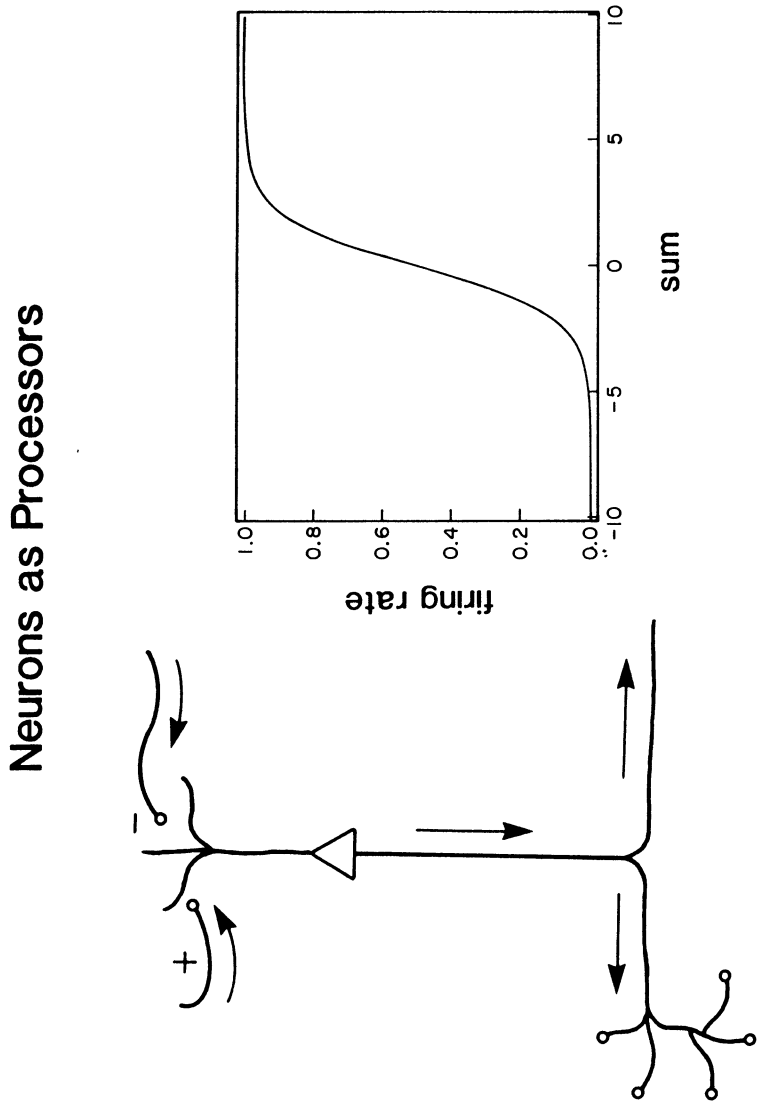
Neurons as Processors
firing rate
1.0
0.8
0.6
0.4
0.2
0.0
-10
-5
0
5
10
sum
1
+

Fig. 1

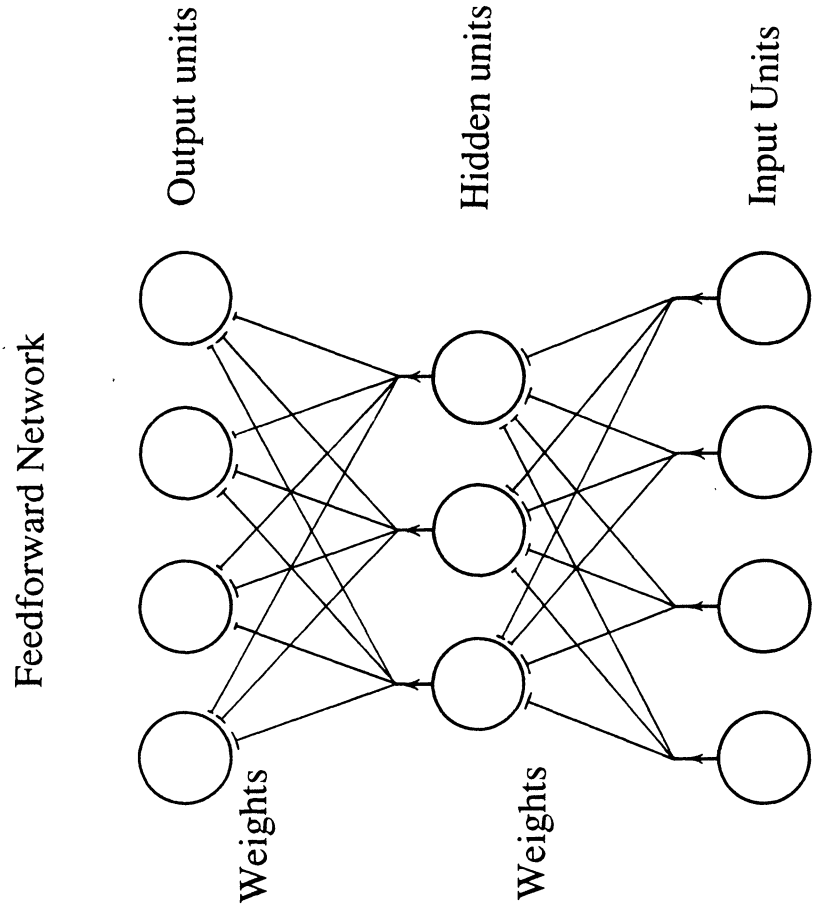
Feedforward Network
Output units
Hidden units
Input Units
Weights
Weights

Fig. 2

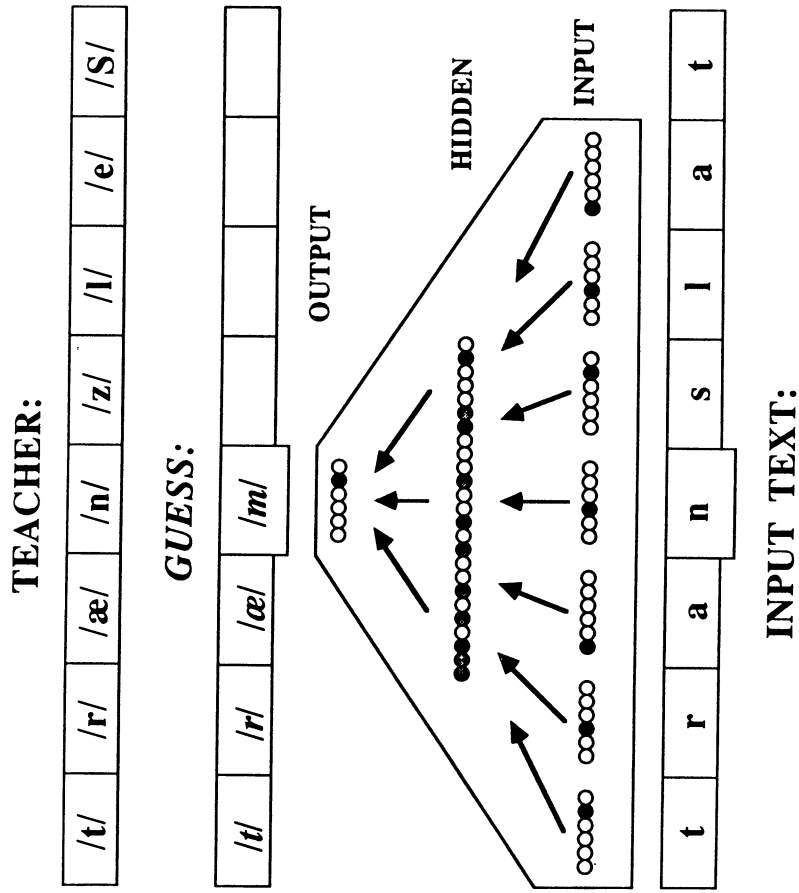
TEACHER:
/t/
/r/
/æ/
/n/
/z/
/l/
/e/
/S/
GUESS:
/t/
/r/
/æ/
/m/
OUTPUT
HIDDEN
INPUT
t
r
a
n
s
l
a
t
INPUT TEXT:

Fig. 3

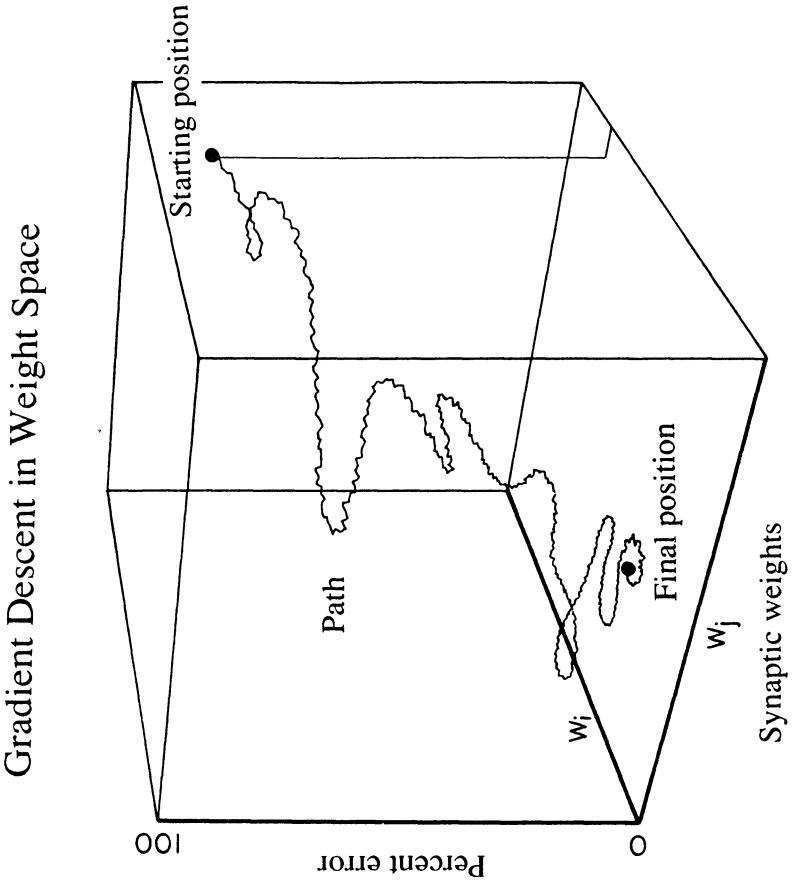
Gradient Descent in Weight Space
Starting position
Path
Final position
$w_i$
$w_j$
Synaptic weights
100
0
Percent error

Fig. 4

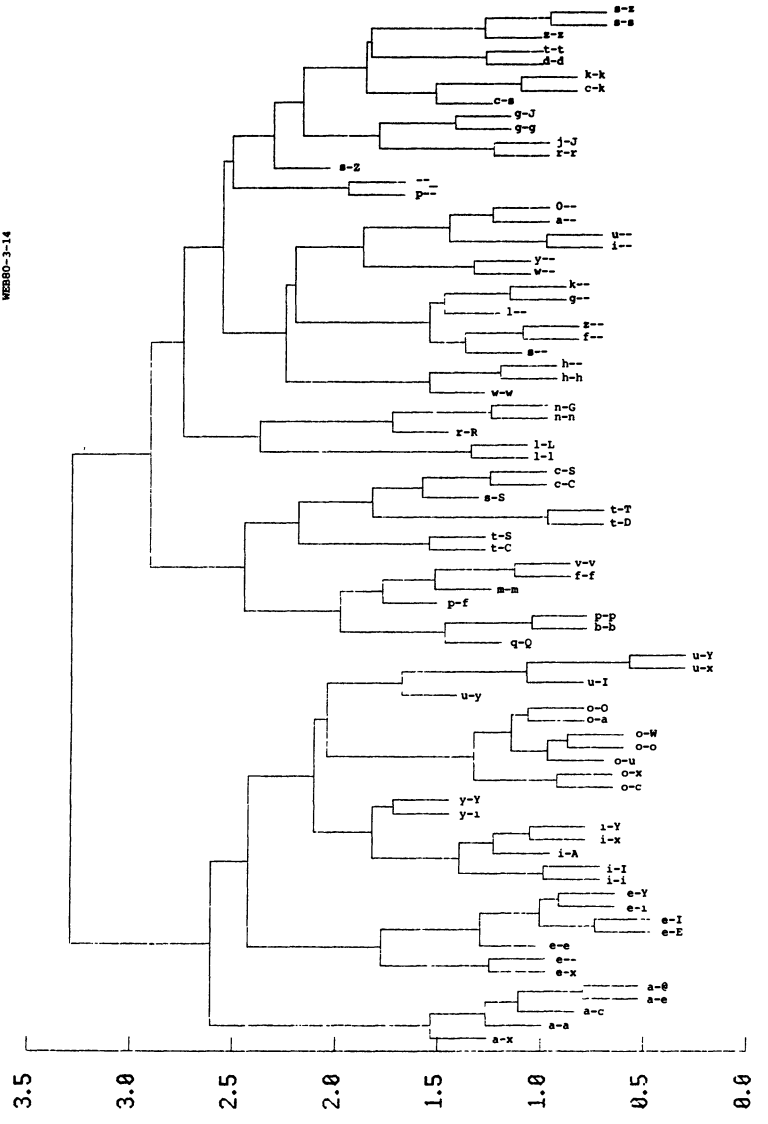
WE3BO-3-14
s-z
s-s
z-z
t-t
d-d
k-k
c-k
c-s
g-J
g-g
j-J
r-r
s-Z
--
p--
0--
a--
u--
i--
y--
w--
k--
g--
l--
z--
f--
s--
h--
h-h
w-w
n-G
n-n
r-R
l-L
l-l
c-S
c-C
s-S
t-T
t-D
t-S
t-C
v-v
f-f
m-m
p-f
p-p
b-b
q-Q
u-Y
u-x
u-I
u-y
o-O
o-a
o-W
o-o
o-u
o-x
o-c
y-Y
y-ı
ı-Y
i-x
i-A
i-I
i-i
e-Y
e-ı
e-I
e-E
e-e
e--
e-x
a-@
a-e
a-c
a-a
a-x
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Fig. 5

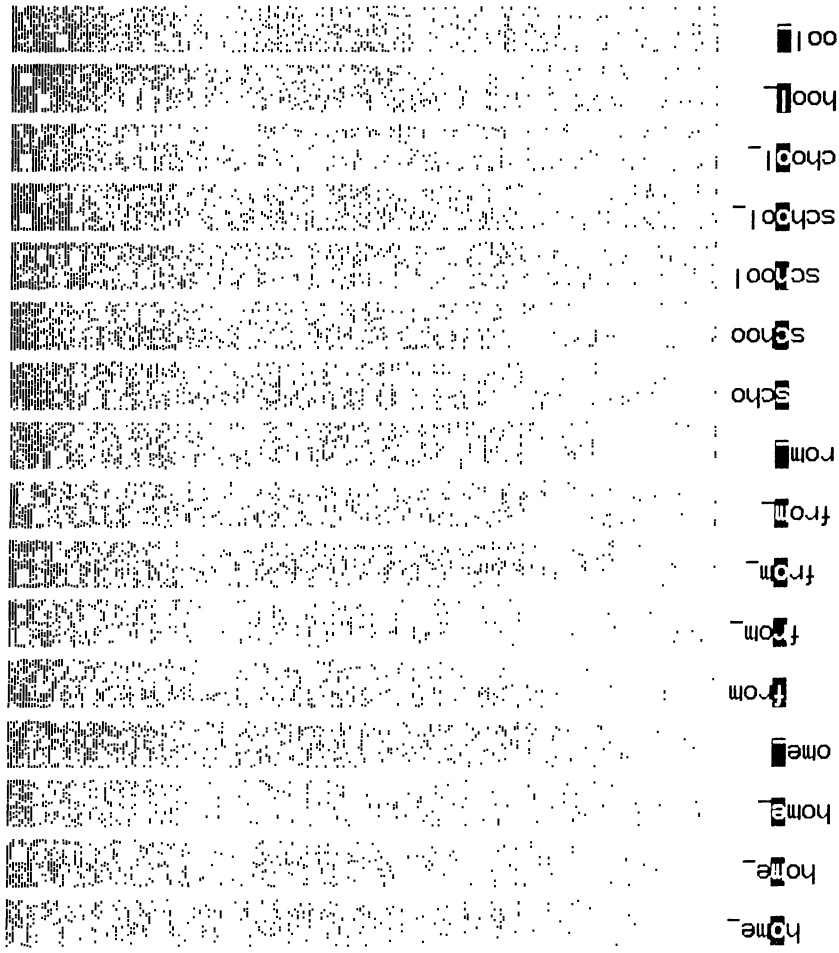
home_
home_
home_
from
from_
from_
from_
from_
rom_
scho
scho
scho
scho
scho
scho
scho

Fig. 6

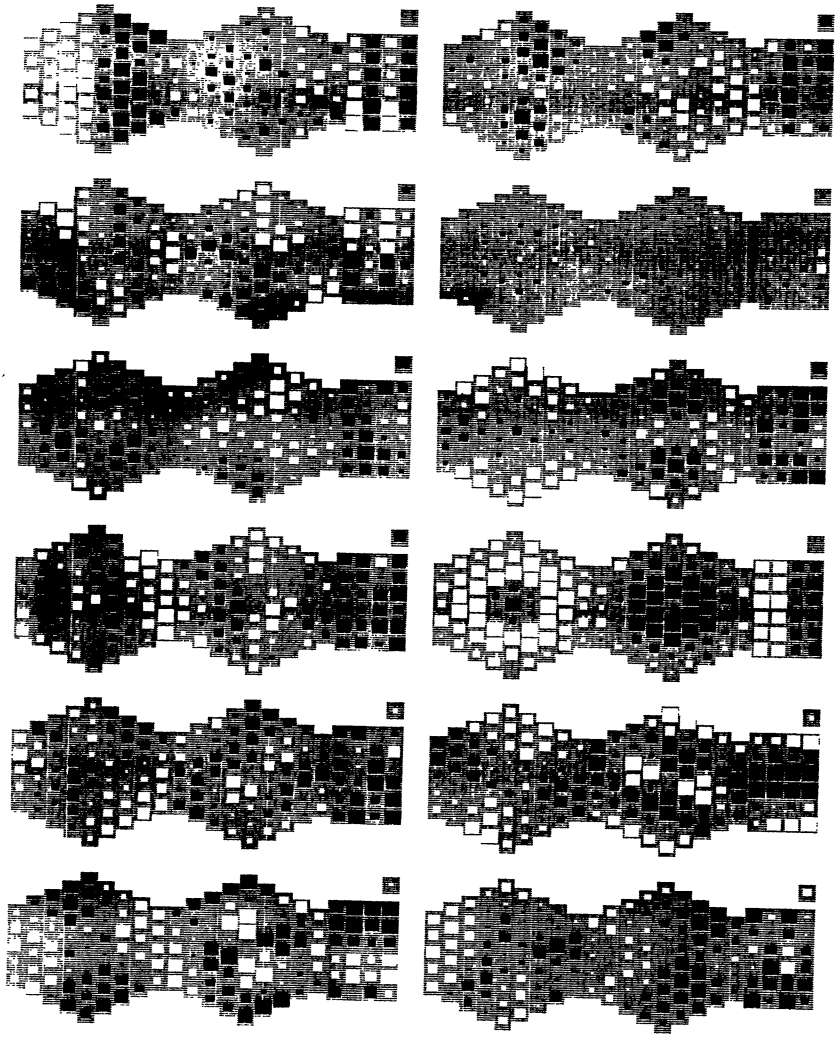

Fig. 7