[Place, U. T. (1989). Towards a connectionist version of the causal theory of reference. *Acta Analytica*, *4*(5), 71-97.
As published in 1989, with minor modifications, May 1993]

TOWARDS A CONNECTIONIST VERSION OF THE CAUSAL THEORY OF REFERENCE[i]

ULLIN T. PLACE

The serial-digital computer as a model of human intelligence.

Until comparatively recently, theories of artificial intelligence were invariably constructed on the assumption that what has come to be known as "the mind-brain" functions in much the same way as the kind of serial-digital computer which, in recent years, has increasingly come to dominate all aspects of contemporary life. The assumption was that the mind-brain solves the problems of adjustment confronting the organism by computing the answers to precisely stated questions in a step-by-step fashion in accordance with a computer program which determines precisely what happens at each step in the light of the outcome of the immediately preceding step. The power of this kind of serial digital computer depends, not on any kind of intuitive grasp of complex issues, but rather on the speed and efficiency with which it can carry out what for a human being is the boring repetitive task of searching through a long list of alternative possibilities until it finds one or more items which fit the stipulated requirements.

There are a number of reasons for thinking that the serial-digital computer with which we are all familiar is not, in fact, a very good model for the functioning of the human or, for that matter, the animal brain:

(1) the serial-digital computer is designed to carry out quickly and efficiently computational tasks which the human brain performs slowly and inefficiently, if at all;

(2) trained human and animal intelligence is characterised by its intuitive grasp of complex issues, such as those involved in visual space perception, which, as far as we know, does not depend on any kind of searching through lists of

alternative possibilities; it is also much quicker and more efficient in performance of such tasks than a device, however powerful, which has to rely on this kind of systematic searching through lists of alternatives;

(3) the time taken by the activity in one neuron in the brain to excite another neuron adjacent to it is much too long for the brain to be able to run through the number of sequential steps it would need to run through in order to compute the solution to the kinds of problem it is able to solve in the time it takes to solve them, if it did in fact operate in the step-by-step manner that a serial-digital computer operates;

(4) the model of the brain as a serial-digital computer requires that data (information) be stored in one or more spatially located stores from which the data is retrieved as and when stipulated by the controlling program. No evidence for the existence of such a localised memory store in the brain has ever been forthcoming from studies of the way in which brain functioning is affected by lesions at different sites within the brain. Phenomena like retrograde amnesia in which loss of memory for past events as a consequence of brain injury or damage is greatest for the most recent events in the individual's past history, with progressively less effect the further back in time the recollection extends and the more often the event in question has been recollected in the past, make it tolerably certain that the individual's ability to remember both facts and past events is a matter of "stamping in" connections widely distributed through the brain, rather than storing information in a localised memory store.

The connectionist model of brain functioning.

The term "connectionism"[ii] has recently been introduced to describe a theory of artificial intelligence which proposes that the correct model for understanding the way the mind-brain functions is not the serial digital computer with which we are all familiar, but rather the device known as "a parallel distributed processor" or PDP. Although the term "parallel distributed processor" was not then used to describe them, the first devices of this kind were constructed more than thirty years ago in the very early stages of research in artificial intelligence before the serial digital computer had come into its own. At this time in the 1950's the object of the exercise was to construct an electronic device in which valves or later transistors are wired up - in the way that the neurons of the brain are wired up - in the form of a network through which a pattern of excitation is transmitted from input to output. Each unit fires or fails to fire depending on the input it receives or fails to receive from two more units behind it in the network. If the neuron in question fires as a result of this excitation it will in turn contribute either to the excitation or to the inhibition of two or more units in front of it in the network.

Now let us suppose that the properties of the units in such a network are such that each time the activity in a particular unit is excited or inhibited by the output from units anterior to it in the network, its susceptibility to that effect in the future is enhanced. It now turns out that a network of units arranged in this way displays properties which are remarkably like those of a living organism whose behaviour is controlled by a brain. As long ago as the 1940's and 1950's such devices were constructed simply in order to see how far systems of this kind could reproduce simple behavioural functions like those of classical conditioning as studied by Pavlov. Such early studies were rapidly overtaken from the 1960's by the development of the serial-digital computer model of brain functioning.

The serial digital computer model of the functioning of the brain superseded the early neural network models because it seemed to offer a way of accounting for much more sophisticated mental processes than anything of which the neural networks were capable. But despite some impressive achievements, workers in artificial intelligence have become increasingly aware in recent years of its limitations. The recent revival of interest in models based on the neural network principle has been motivated by the serious difficulties which were encountered in programming a serial digital computer to perform what for a human being or an animal are relatively simple sensory discrimination or pattern recognition tasks, an ability which computers must acquire if robots are to take over the kind of routine inspection tasks currently performed by human operators. It turns out that if a parallel distributor processor is harnessed to the appropriate sensors, it can rapidly learn to recognise complex patterns of sensory information without having to check and without having to be specifically programmed to check a list of alternative possibilities in the way the serial digital computer does. It can do this, moreover, without being defeated, in the way the serial digital computer invariably is, by a familiar pattern presented in a way which the system has never previously encountered and with which it has not been specifically programmed to deal, in other words, problems such as that presented by a familiar scene viewed from an unfamiliar vantage point, or at a very different season of the year from that which obtained when it was previously encountered.

From the point of view of psychology and the neuro-sciences, the importance of the parallel distributed processor is that we now have a much better model for the way the brain actually functions than that provided by the serial digital computer, not only in those areas, such as pattern perception, where the PDP replicates the human and animal ability to discriminate complex patterns in a variety of novel contexts in a way that the serial computer model is unable to do, but also in areas, such as learning the past tenses of English verbs (McClelland, Rumelhart and the PDP Research Group, 1986, Vol 2, p.216 ff.), where, unlike the serial computer, the PDP makes exactly the same kind of errors that children make in learning the same task.

What is perhaps more difficult to defend is the claim made by some philosophers, Patricia Churchland (1986), Paul Churchland (1988) for example, as well as by some computer scientists, such as Paul Smolensky (1988), that connectionism also has profound implications for the philosophy of mind. The basis of this claim is undoubtedly the very considerable vested interest which philosophers, particularly in the United States, have acquired in recent years in the project of artificial intelligence and cognitive science. I am thinking in particular of the work of philosophers like Hilary Putnam, Jerry Fodor, Dan Dennett, John Searle, Stephen Stich and the Churchlands themselves. I am personally inclined to think that this vested interest has been acquired on false pretences, and that scientific research in this area is being hampered by the philosopher's insistence on importing into the field of cognitive science, philosophical muddles about such issues as intentionality and the causal theory of reference which, in my view, the field can well do without.

Be that as it may, the vested interest exists; as does a vested interest on the part of <u>some</u> of the philosophers I have mentioned - of whom Jerry Fodor is the most striking example - in the serial computer model of the functioning of the mind-brain. Moreover, it is in the light of this vested interest in the serial-digital computer model that we need to understand Fodor and Pylyshyn's (1988) recent attack on connectionist theories in Cognition which has been ably reviewed and rebutted by Steve Mills in a recent paper published in this journal (Mills 1989). Steve's paper deals with Fodor and Pylyshyn's attempts to get round the difficulty presented to the serial-digital computer model by the objection that neural activity proceeds at much too slow a rate for it to run through all the steps which it would need to go through if it were organised like a serial-digital computer. What Steve does not attempt to address is Fodor and Pylyshyn's other argument.

Although Fodor and Pylyshyn end up by conceding that neural networks may "sustain some cognitive processes", in particular "such processes as can be analyzed as the drawing of statistical inferences" (Fodor and Pylyshin 1988, p.68), they reject the claim made by Rumelhart and McClelland (Rumelhart, McClelland and the PDP Group 1986, p.110) that "PDP models could form a reasonable basis for modeling cognitive processes in general." The grounds for this rejection are that in the area of language, logic and symbolic representation in general, where the serial digital computer model has a powerful and proven theory, the PDP model has nothing to contribute. It is this second part of Fodor and Pylyshyn's critique of connectionism that I want to address in this paper.

In order to show that a connectionist account of language, logic and symbolic representation is not only possible, but, in so far as its shape can be envisaged, is actually superior to the more highly developed theory based on the model of the serial-digital computer, I have decided to approach the issue by presenting a connectionist version of the causal theory of reference. I have three reasons for this choice:

(1) the aspect of language which lends itself most readily to an explanation in terms of the operation of a parallel distributed processor is the process whereby a child that is learning its first language acquires the ability to make indexical reference to a particular or a kind, in a case where the particular itself or an instance of the kind is present in the common stimulus environment of speaker and listener;

(2) for reasons which I don't altogether understand, and which in any case would take us too far afield to explore, the causal theory of reference in the form of Putnam's (1975) Twin Earth example has become a key issue in Fodor's functionalist theory, as illustrated by his recent book Psychosemantics (Fodor 1987) which in turn is deeply committed to the serial-digital computer as a model for the functioning of the mind-brain;

(3) for the past four years,[iii] I have been wrestling with the idea that there is something deeply implausible about the causal theory of reference, as interpreted by Kripke (1972;1980) and Putnam (1975) in terms of the theory of rigid designation, particularly when it is viewed from the standpoint of a psychologist who is concerned to explain how a listener can learn to identify the particular or kind to which a speaker is referring; since the strength of connectionism lies in the accuracy with which it reproduces both the microstructure and the function of the brains of

living organisms, it follows that, if a successful connectionist theory of reference can be developed, it should have a high degree of "face validity" as a plausible causal psychological account of how this function is actually performed by the human brain.

The causal theory of reference and the traditional view

What, then, is the causal theory of reference? The standard way of introducing the theory and the way that I propose to adopt, is to contrast it with the more traditional theory of reference which holds with the Port Royal logicians (Arnauld and Nicole 1662) that the "extension" of a general term, in other words, the class of actually existing objects to which the term applies, is determined by what they called its "comprehension", but which is now generally referred to, following Sir William Hamilton (Kneale & Kneale, 1962 p. 318) as its "intension" spelt-with-an-s. If, like me, you find it helpful to think about such things in terms of a simple mechanical analogy, you can think of the intension of a general term as a kind of electronic filter which allows those objects which satisfy the relevant criteria to pass through the filter and become members of the class of actually existing objects which constitute the term's extension, while excluding others which fail to satisfy the criteria.

What is essentially the same theory as applied to singular terms is to be found in Frege's (1892) doctrine that the "reference" or Bedeutung of a singular term, i.e., the one and only actually existing individual to which the term refers, is determined by its "sense" or Sinn. Here again we can think of the sense of a singular term, like the definite description The man *we met in the pub last night*, as a filter which, in this case, picks out a single unique individual.

In contrast to this the standard version of the causal theory of reference holds that, at least in the case of proper names and natural kind terms, it is the other way round: the intension or sense of a linguistic expression is determined by its extension or reference, or, more accurately, in the case of proper names and natural kind terms, the linguistic expression "rigidly designates" (Kripke 1972;1980) its extension or referent without any mediating intension or sense.

The connectionist version of the causal theory

What I am calling "the connectionist version of the causal theory of reference", on the other hand, retains the traditional view that in all cases extension and reference are determined by intension and sense. Since it retains the traditional view that the direction of causal action is from intension and sense towards extension and reference, rather than *vice versa*, it might be argued that it is not really a version of the causal theory, as ordinarily understood. I would claim, nevertheless, that there is some justification for that description, not only because the theory uses connectionist principles in order to accommodate features of the reference of proper names and natural kind terms which give plausibility to the more orthodox versions of the causal theory, but also because, in the case of natural kind terms, it involves an explanation of how what we may call "the natural extension" of the kind in question acts so as to bring the intension of the term into line with that extension.

The problem of natural kind terms for the traditional theory

The problem which *natural kind terms* present for the traditional theory according to which the extension of such terms is determined by their intension or sense is this. Suppose we define the intension of a general term as the set of criteria that we use in assigning an instance to the class of objects which constitute the extension of that term. It now turns out that in the case of natural kind terms, terms, that is, which denote such things as a biological species or a naturally occurring substance or stuff, we cannot, at the level of ordinary language, specify any set of criteria that will

(a) include all members of the recognised extension of the term without including some that we would not normally and naturally include, or

(b) exclude all instances which are not members of the recognised extension of the term without excluding some instances that we would not normally and naturally exclude.

It is only after scientific research has revealed the so-called "real essence" of the natural kind in question that it becomes possible to provide a precise set of criteria of this kind.

Yet how do we explain the fact that long before the chemical composition of water was discovered, human beings were reliably classifying instances of what we now recognise as cases of H_2O and distinguishing them from liquids with different chemical compositions?

The explanation that is offered by the orthodox versions of the causal theory of reference is that without knowing the real essence of the natural kind in question (H_2O in the case of water), pre-scientific language users nevertheless have a kind of intuitive grasp of the abstract object constituted by a natural kind such as water, without the need for any criteria for deciding whether a particular instance is an instance of the kind in question. Natural kind terms, that is to say, "rigidly designate" the same extension as is subsequently picked out by the intension of an expression such as H_2O which specifies the real essence of the natural kind in question, as and when this is discovered by empirical research.

The explanation that is offered by what I am calling "the connectionist version of the causal theory of reference" is based on the supposition that the intension or sense of an expression, both in the case of the individual and within the linguistic community as a whole, is subject to continuous slow modification as a result of the shaping of linguistic usage by the experience of success and failure in establishing reference and determining appropriate and practically useful extensions. What I am suggesting is that the way we classify the features of our environment, both as individuals and collectively as a linguistic community, is modified by the success and failure of different ways of classifying things in helping us to understand and control our environment. Classifications of which no instances are found or which ignore important distinctions are abandoned in favour of those which separate out features which are of practical and theoretical importance. As a result of this process, natural kind terms like "water", as used in ordinary language, gradually acquire an extension which coincides with that of a technical expression such as "H_2O in its liquid form" which describes what is later discovered to be the real essence of the natural kind in question. In many cases, moreover, it does this long before the real essence is empirically discovered by scientific research.

The problem of proper names for the traditional theory

The problem which the Fregean theory of reference encounters in the case of proper names is well known. The problem is that *any* description that is true of the bearer of a proper name can under appropriate circumstances be used to explain which individual the name is being used to refer to. For example we can identify the individual that the proper name Julius Caesar refers to by mentioning one or more of the following facts about that gentleman's historical career:

(a) The Roman general who conquered Gaul

(b) The author of the De Bello Gallico

(c) The Roman general who invaded Britain in 55 and 54 B.C.

(d) Pompey's main rival

(e) The Roman general who crossed the Rubicon with his army in 50 B.C.

(f) The man who was murdered in the Roman Senate on the Ides of March 44 B.C., etc., etc., etc.

Yet we cannot say that any one or any collection of these descriptions constitutes the sense of the proper name. For one thing, the name was in use and understood as referring to the individual in question long before any of the events mentioned in these descriptions took place. For another, two different people can identify the bearer of the same proper name by two quite different and non-overlapping sets of descriptions as in Frege's (1918) example of Dr. Gustav Lauben who is known to Herbert Garner, but to no one else, as the man who "was born on 13th September, 1875 in N.N.", but who "does not know where Dr. Lauben now lives nor indeed anything about him"; whereas Leo Peter who knows Dr. Lauben both personally and as "the doctor who lives as the only doctor in a [certain] house .... does not know that Dr. Lauben was born on 13th September, 1875 in N.N."

In this case the suggestion which is made by orthodox versions of the causal theory of reference is that proper names simply refer to their bearers without any mediating sense. This seems eminently plausible. Here, surely, we have the perfect case of a rigid designator.

Yet doubts remain. The psychological processes whereby we recognise someone or something in our immediate stimulus environment as the bearer of a proper name, when that name is simultaneously uttered, and understand a reference to the bearer when the bearer of the proper name is absent, remains totally mysterious on this view. Moreover, it is precisely this psychological process that what I am calling "the connectionist version of the causal theory of reference" as applied to proper names aims to elucidate.

As applied to the case of proper names, the connectionist version of the theory begins by drawing a distinction between

(a) the intension or sense of an expression as it is understood by a particular individual, and

(b) the intension or sense of an expression as it is understood within the wider linguistic community.

In the case of the intension or sense of an expression as it is understood within the wider linguistic community, all we can say about the intension or sense of a proper name is that, apart from some indications as to the kind of object to which the name refers which are given by the form and context of the utterance within which the name occurs, its sense restricts its reference to that single individual to which the name, when used in that sense, was originally assigned when it was first given.

The problem with this account is that the only thing that distinguishes the sense of one proper name from that of another is the date and place at which the proper name in question was first assigned to the individual in question. This makes some kind of sense in those cases where a proper name is assigned by some recorded ceremony such as baptism in the case of a human infant; but even in this case, as Frege's example of Dr. Lauben shows, there is no requirement that those who use the name correctly should know when and where it was first used as the name of that individual. In many cases, moreover, place-names of ancient origin, for example, no record of any such ceremony exists, if indeed it ever took place.

It follows that proper names can only function in language as they do in so far as each proper name whose reference is understood by a particular listener has a distinctive intension or sense for that particular listener which determines its reference to the individual in question. This intension or sense which is unique to the individual consists in that individual's ability to identify the bearer of the name, either by its visual appearance, characteristic sound or feel, or by some descriptive predicate that is true or generally held to be true of it. In cases where the bearer of a proper name is widely known, either by visual appearance or by description (usually both), throughout the linguistic community, the proper name can be said to have an intension or sense within the wider linguistic community which approximates to that of a general term. The proper name Margaret Thatcher is a case in point.

Derivation of the theory from the principles of connectionism

The most striking characteristic of a parallel distributed processor is that it can learn to recognise stimulus patterns corresponding to

(a) the same individual presented in widely differing guises,

(b) instances of the same kind of thing which differ amongst themselves in other respects, provided that

(a) it has had previous experience of a sufficiently wide range of both positive instances (Skinner's S^D) and otherwise similar negative instances (Skinner's S^{Δ}),[iv] and

(b) it is given reliable feed-back as to whether its judgement in a given case is correct (Skinner's "reinforcement") or incorrect ("disinforcement"[v]).

It is not difficult to see from this that the hypothesis that the brain functions as a parallel distributed processor can readily account for those cases where the listener's grasp of the intension or sense of a linguistic expression consists in the ability to pick out the individual in question, where the referent is an individual, or an instance of the kind, where the referent is a kind, provided that the listener is confronted by that individual or instance or by some kind of representation of it, such as a photograph or drawing. It can also readily account for the process whereby the boundaries between the intension of one expression and that of another are shifted, in the case of the individual language user, as a result of corrections by and failures of comprehension on the part of the listener, and, in the case of the linguistic community as a whole, as a result of some practical utility which is achieved by shifting the previously accepted conceptual boundaries in this way.

The problem of absent instances

What is less readily accounted for on connectionist principles is the listener's ability to grasp the intension or sense of an identifying reference to an individual or a kind, when neither the individual nor an instance of the kind are present in the listener's current stimulus environment. But it is far from clear that any rival theory is any better placed in this respect. Moreover, there are some cases of reference to individuals and kinds which are absent from the context of utterance where it seems reasonable to propose that at least part of what is involved in grasping the intension or sense of an identifying reference in such cases consists in the truth of a counterfactual to the effect that one would be able to pick out the individual or instance of the kind, if they were part of one's current stimulus environment.

However, another and perhaps more important part of what is involved in grasping the intension or sense of expressions referring to individuals and kinds which are absent from the listener's current stimulus environment is the listener's ability to construct, assert and act on sentences which contain singular terms whose referent is not and never has been part of his or her stimulus environment whether "in the flesh" or in the form of a pictorial representation or audible record, or which contain general terms of which no instance is or ever has been part the listener's stimulus environment. In order to provide a connectionist account of what is involved in the kind of knowledge which speaker and listener must share for the listener to grasp the speaker's identifying reference to an individual or kind which has never been part of the listener's stimulus environment, we need a connectionist account of

(a) sentence construction and interpretation,

(b) the way sentences map onto the reality they depict,

(c) what makes such sentences true or false, and

(d) how the truth and falsity of such sentences is discriminated by the listener.

A connectionist approach to syntax, semantics and epistemology

On the face of it, the serial computer model of brain functioning would appear to have a head start over the parallel distributed processor in this area, especially as far as the theory of sentence construction is concerned. Nevertheless, there are a number of considerations which suggest, not only that we shall not have to wait very long before a connectionist theory of these phenomena becomes available, but also that, when it does, it will in fact model the performance of linguistically competent human beings very much more closely than any serial computer model can hope to do.

The normative character of syntactic and logical principles

In the first place there is the observation that the rules of sentence construction in the serial computer model are programmed in such a way that they become causal laws governing the behaviour of the system. Although these laws may be overridden by other laws which tend in a different and opposite direction, their manner of operation is quite different from that of the principles of syntax and logic as they apply to human language and thought. For the laws of syntax and logic are not laws of nature governing the production of language and thought. They are normative principles in the light of which language and thought is criticised *after* its production.

This is essentially the same point that was made by Frege when he criticised the "psychologism" of Husserl's (1891) book *Philosophie der Arithmetik*. Psychologism, as Frege presents it in his review of Husserl's book (Frege 1894), is the mistake of treating the principles of logic as if they were empirically discovered causal laws governing the process of human thought. However, Frege's point is not, as is sometimes supposed by those who seek to defend computational theories from the charge of fallaciously confusing reasons with causes, that it is part of the nature of logical principles that they *cannot* function as causal laws. Fodor (1987) is clearly right to point out that the causal role which logical principles play in computer software is sufficient refutation of that claim. Frege's point is that, whatever may be true of a serial computer (not that he was familiar with such devices), <u>that</u> is not how logical principles function in the regulation of human thought. The role of logical principles in relation to human thought is essentially normative. They tell us how we ought to think, not how we actually think. Moreover, it is part of the concept of a normative principle that its function is to provide a standard to which, left to its own devices and obeying its own intrinsic laws, human behaviour and mental processes would not conform, but to which they are by and large induced to conform by the favourable social consequences of so doing and the unfavourable social consequences of failing to do so. In the case of the principles of syntax, the penalty for failure to conform is failure to communicate. In the case of the principles of logic, it is the failure to convince. It is evident, moreover, that learning to conform to syntactic and logical principles in the light of indications of success and failure, provided by the response or lack of response on the part of the listener, is just the kind of skill which a parallel distributed processor is well adapted to acquire. The serial computer model, by contrast, requires a complex step-by-step program in order to construct well formed sentences in natural language. Such a program is much easier to envisage as "hardwired" or genetically preprogrammed, than as acquired by the ordinary processes of learning. But whether it is thought of as learned or innate, in so far as syntactic and logical principles are conceived as generating thought and language, rather than as correcting what has already been generated, this theory leaves no room for the kind of mistakes which consist in a failure to conform to syntactic and logical principles, and which are only too common in human thought and language production.

A connectionist model, on the other hand, would see the process whereby one word, phrase, sentence or thought leads to another, as proceeding according [to] principles like the classical Laws of Association, modulated by a feed-back mechanism which by and large ensures the conformity of the thought process to the standards of syntactic and semantic coherence, logical validity, truth and relevance which are required in order to secure indications of successful communication from the listener. Such a model has no difficulty in handling the occasional failure of communication which results from failure to conform to the principles in question. Such failures are, in any case, an essential part of the process of trial-and-error whereby, on a connectionist view, conformity to such principles is learned in the first place.

Connectionism and semantics

We thus have reason to think that, when the connectionist story of how the ability to conform to the principles of syntax and logic is acquired by the brain comes to be told, it will be very much closer to the way the brain actually works than anything the serial computer model can now provide or is likely to provide in the future. Nevertheless, it has to be conceded that existing serial computer models of syntactic and logical competence are far in advance of anything that can currently be proposed in terms of the model of the parallel distributed processor. But this advantage only applies to the case of syntax and logic. When it comes to giving an account of the semantic aspects of linguistic competence, of the listener's ability to understand what is said, the best that the serial computer model can currently come up with is the bizarre theory proposed by Katz and Fodor (1963) in which comprehension is interpreted as a matter of retrieving, from an in-built lexicon in the brain, an entry - written, presumably in Fodor's (1975) private language of thought - corresponding to each word in the sentence as it appears. Although he makes no mention of the Katz and Fodor theory - presumably because he thinks its absurdity is too obvious - it is the lack of any coherent account of comprehension which is at the root of Searle's (1980) Chinese Room argument against taking the serial computer as a model for the way the brain functions in this regard.

But where the serial computer model falters, the parallel distributed processor comes into its own. For, as we have already seen, it is not difficult to provide a convincing connectionist theory of the listener's grasp of the intension or sense of a referring expression in those cases where the referent, or an instance of it where the referent is a kind, is present in the listener's stimulus environment. The only problem for the connectionist view is to account for what is referred to in those cases where there is no instance or representation of the referent in the stimulus environment and where the listener's grasp of reference cannot be a matter of the listener's ability to pick out the referent or an instance of it, were it or a representation of it to appear.

A case in point is one where we understand a reference to Homer without having the slightest idea what he looked like or how his voice sounded. In such a case, we understand the reference to Homer if and only if we understand a sentence of the form "Homer is the author of the _Iliad_ and the _Odyssey_" and know it to be true. Moreover, in order to understand _that_ sentence, we need to be able to understand and assent to some such sentence as "The

Iliad and the Odyssey are two very long poems written in Archaic Greek, the one telling the story of the Trojan War, the other the story of the travels and adventures of Odysseus on his return journey from Troy to his home in Ithaca." It follows from this that in order to explain the listener's grasp of the reference of the name Homer and that of the names of persons and places referred to in his poems, we need to be able to explain the listener's grasp of the intension or sense of complete sentences, particularly sentences which contain reference to objects, events and states of affairs which are either fictitious or so far removed in space and time from the context of utterance as to be inexplicable, in any direct way, in terms of the listener's ability to pick out features of his or her actual or possible stimulus environment.

There is, no doubt, an expectation in some quarters that the serial computer model will eventually be in a position to provide an adequate semantics for sentences by incorporating a version of formal truth-conditional semantics in the tradition that goes back to Tarski (1956). But without entering into the thorny topic of the adequacy of such an approach, its obvious affinity with that of formalised syntax and logic leads to the expectation that, however successful the project of writing a serial computer programme for truth conditional semantics may turn out to be, it will tell us more about how to replicate the comprehension of a sentence in a serial computer, than about how sentence comprehension is actually achieved by the human brain.

Towards a connectionist theory of sentential semantics

This, needless to say is neither the time nor the place to develop a systematic connectionist theory of the semantics of sentences.

Nevertheless, it is not difficult to provide some pointers in that direction. We may begin, perhaps, by noting that one of the most striking properties of a parallel distributed processor is that if it is presented with three distinct input types, A, B, and C, such that C occurs if and only if it has been immediately preceded by the combination AB, the system will rapidly learn to expect C, given the combination AB, and not to expect C, given A and B on their own.

We can now state the problem that is presented for such a system by the semantics of a simple sentence like _The cat is on the mat_. Any listener who understands the words _cat_ and _mat_ and who is familiar with sentences of the form _The X is on the Y_ will know what to expect when he or she hears that sentence, despite the fact that he or she may never before have encountered

(a) that particular combination of words,

(b) a cat on a mat, or

(c) that particular combination of words, immediately followed by seeing a cat on a mat.

We may assume, however, that in the process of learning the meaning of the constituent words of which the sentence is composed, the listener has repeatedly encountered cases where

(a) hearing the word _cat_ has been accompanied by seeing a cat or a picture of a cat,

(b) hearing the word _mat_ has been accompanied by seeing a mat or a picture of one, and

(c) hearing sentences of the form _The X is on the Y_ has been accompanied by seeing the first object on top of the second.

Given that assumption, all that is needed to explain the ability of the sentence _The cat is on the mat_ to create in the listener the expectation of seeing a cat on a mat, despite the fact that he or she has never encountered such a phenomenon before, is some account of the process whereby the listener learns to combine the expectations aroused by each of the three experience-based expectations into a single expectation which, as such, has no basis in previous experience.

Since there would seem to be no great problem in envisaging how a PDP system might be supposed to combine its expectations in this way, we may take it that the problem of providing a connectionist theory of the semantics of simple sentences like The cat is on the mat is not far from solution. But if it can be done in the simple case, there would seem to be no reason in principle why the same pattern of explanation should not eventually be extended to cover all cases, even the most complex and difficult.

NOTES

i. A revised version of a paper presented in the Department of Philosophy, University of Zagreb, on the 14th of November 1988.

ii. The standard textbook on connectionism is the two volume work by Rumelhart, McClelland and the PDP Research Group (1986).

iii. Since September 1986 when I attended the first of three courses held in successive years (1986, 1987 and 1988) at the Inter-University Graduate Centre, Dubrovnik, all three of which were directly or indirectly concerned with Kripke's (1972;1980) theory that proper names and natural kind terms are 'rigid designators', and with the roots of that doctrine in formal modal logic. I am indebted to many of my fellow participants in those courses, and to David Charles and Nathan Salmon in particular, for such understanding of these matters as I possess. I should emphasise, however, that I alone am responsible for any defects of understanding as may appear in the way these matters are treated here.

iv. The reference is to the account of operant discrimination learning given by B. F. Skinner in Chapter Five of his 1938 book, Behavior of Organisms. The parallel between Skinner's account of discrimination learning in organisms such as the rat and the pigeon and the way a PDP learns to discriminate patterns in a sensory array is quite remarkable.

v. To use the term proposed by Harzem and Miles (1978).

REFERENCES

Arnauld, A. and Nicole, P. (1662) La Logique, ou l'art de penser. Paris.

Churchland, P. M. (1988) Matter and Consciousness, 2nd ed. Cambridge, Mass.: M.I.T.Press.

Churchland, P. S. (1986) Neurophilosophy: Toward a Unified Science of the Mind/Brain. Cambridge, Mass.: M.I.T.Press.

Fodor, J. A. (1987) Psychosemantics. Cambridge, Mass.: M.I.T. Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. Cognition 28: 3-71.

Frege, G. (1892) Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik, 100: 25-50. English translation as 'On sense and reference' by M. Black. In P. T. Geach and M. Black (eds.) Translations from the Philosophical Writings of Gottlob Frege. Oxford: Blackwell, 1952.

Frege, G. (1894) Review of E. G. Husserl Philosophie der Arithmetik. Zeitschrift für Philosophie und philosophische Kritik, 103: 313-332. English translation of "illustrative extracts" by P. T. Geach. In P. T. Geach and M. Black (eds.) Translations from the Philosophical Writings of Gottlob Frege. Oxford: Blackwell, 1952.

Frege, G. (1918) Der Gedanke. Eine logische Untersuchung. Beiträge zur Philosophie des deutschen Idealismus I: 58-77. English translation as 'The thought. A logical enquiry' by A. and M. Quinton. In Mind (1956) LXV: 289-311.

Harzem, P. and Miles, T. R. (1978) Conceptual Issues in Operant Psychology. New York: Wiley.

Husserl, E. G. (1891) Philosophie der Arithmetik. Halle: Saale.

Katz, J. J. and Fodor, J.A. (1963) The structure of a semantic theory. Language, 39: 170-210.

Kneale, W. and Kneale, M. (1962) The Development of Logic. Oxford: Clarendon Press.

Kripke, S. (1972) Naming and necessity. In G. Harman and D. Davidson (eds.) *Semantics of Natural Language*. Dordrecht: Reidel.

Kripke, S. (1980) *Naming and necessity*. Dordrecht: Reidel.

McClelland, J.L, Rumelhart, D.E. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*. Cambridge, Mass.: M.I.T.Press.

Mills, S. T. (1989) Connectionism, the classical theory of cognition and the hundred step constraint. *Acta Analytica: Philosophy and Psychology*, 4: 5-38.

Putnam, H. (1975) The meaning of 'meaning'. In K. Gunderson (Ed.) *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis: University of Minnesota Press.

Rumelhart, D. E., McClelland, J. L and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. Cambridge, Mass.: M.I.T.Press.

Searle, J. (1980) Minds, brains and programs. *Behavioral and Brain Sciences*, 3: 417-424.

Skinner, B. F. (1938) *The Behavior of Organisms*. New York: Appleton-Century.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11: 1-23.

Tarski, A. (1956) *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. English translation by J. H. Woodger, Oxford: Clarendon Press.