Rev. Neurol. (Paris), 1994, *150*, 8-9, 570-579.

THE ORGANIZATION OF MEMORY
A PARALLEL DISTRIBUTED PROCESSING PERSPECTIVE

James L. McCLELLAND

SUMMARY

Parallel distributed processing (PDP) provides a contemporary framework for thinking about the nature and organization of perception, memory, language, and thought. In this talk I describe the overall framework briefly and discuss its implications of procedural, semantic, and episodic memory. According to the PDP approach, the processing of information takes place through the interaction of a large number of simple processing units organized into modules. Storage occurs through the modification of connection weights based on the system's response to its input, that provides an opportunity for incremental storage. I will describe how connection modification may give rise through the course of experience to procedural learning and to the formation of semantic memories, structured by their semantic content. I will argue that the discovery of semantic structure requires gradual learning, with repeated exposure to representative samples of the structure to be learned. I will then describe two neuropsychological implications of the PDP approach. First, I will consider the possible modular organization of semantic information in the brain. Then, I will examine the role of the hippocampus in learning and memory. In the first case, we will see how the PDP approach leads us to see how brain damage might produce apparent dissociations between categories, when in fact the underlying organization is not by category but by modality. In the second case, we will see that the PDP approach gives us a new way to understand why it is important that unique, arbitrary associations not be stored all at once in the same memory's systems used for semantic information. This leads to a specific theory of complementary roles of cortex and hippocampus, and to an explanation for the phenomenon of temporally graded retrogade amnesia.

L'organisation de la mémoire. Intérêt du « Parallel Distributed Processing ».

J.L. McCLELLAND. *Rev. Neurol. (Paris)*, 1994, *150* : 8-9, 570-579.

RÉSUMÉ

La méthode dite « parallel distributed processing » (PDP) offre un nouveau cadre à la réflexion sur la nature et l'organisation de la perception de la mémoire, du langage et de la pensée. L'auteur décrit d'abord brièvement l'ensemble de ce cadre et discute ses applications à la mémoire procédurale, sémantique et épisodique. Selon la méthode DPD, le traitement de l'information passe par l'interaction d'un grand nombre d'unités de traitement simple organisées en modules. Le stockage a lieu par modification des poids de connexions, fondée sur la réponse du système à ses afférences, ce qui permet un stockage de plus en plus important. L'auteur décrit ensuite comment la modification des connexions peut aboutir, au cours de l'expérience, à l'apprentissage procédural et à la formation de mémoires sémantiques structurées par leur contenu sémantique. Il démontre que la découverte d'une structure sémantique exige un apprentissage graduel comportant une exposition répétée à des échantillons représentatifs de la structure à apprendre. L'auteur décrit alors deux applications du PDP à la neuropsychologie. Il considère d'abord l'organisation modulaire possible d'une information sémantique dans l'encéphale, puis examine le rôle joué par l'hippocampe dans l'apprentissage et la mémoire. Dans le premier cas, on voit comment la méthode PDP amène à envisager comment une lésion cérébrale peut provoquer des dissociations entre catégories, alors que l'organisation sous-jacente n'est pas fondée sur la catégorie mais sur la modalité. Dans le second cas, l'on voit que la méthode PDP procure une nouvelle façon de comprendre pourquoi il est important que des associations uniques et arbitraires ne soient pas toutes stockées en même temps dans les mêmes systèmes mnésiques que ceux utilisés pour l'information sémantique. Tout cela aboutit à une théorie spécifique des rôles complémentaires que jouent le cortex et l'hippocampe ainsi qu'à une explication du phénomène d'amnésie rétrograde temporellement graduée.

How is memory organized ? This question actually subsumes many more specific questions. To list only two, how many different kinds of memory are there ? What is the substructure of each type ? These kinds of questions have motivated a great deal of research, both psychological and neuropsychological. Behind them lies another set of questions : Why is memory organized as it is ? What leads to its organization ? Are there fundamental computational issues that dictate the form of the human memory system, or is it simply a conglomeration of evolutionary artifacts ?

I will address some of these questions in this talk, drawing on insights from the behavior of connectionist,

Tirés à part : James L. McCLELLAND, Department of Psychology Carnegie Mellon University Pittsburgh, PA 15213, USA.

parallel-distributed processing models of human memory. First, I will briefly review the properties of such models. Then, I will describe how structured procedural and semantic knowledge may arise from learning in such systems. I will argue that both kinds of learning depend on gradual connection weight adjustment through repeated and interleaved presentation to representative samples of the structure of the domains to which they apply.

I will then build upon these descriptions to address two specific questions from the range of issues raised earlier. Each question is at once neuropsychological and psychological. The questions are :

How is semantic memory organized in the brain ? Is it by category or by modality ?

Why is there is a special system – namely the hippocampus – dedicated to the rapid acquisition of episodic memories ?

The investigation of these questions relates closely to the legacy of Charcot, since they depend on the results of the study of neurological patients, and they draw on a consideration of the effects of brain damage on memory. They bring this kind of study together with explicit computational modeling based on the principles of parallel distributed processing (PDP). In both cases, I will argue that the properties of PDP models allow us to supplement the careful analysis of patient behavior with theoretical observations that can dramatically advance our understanding of the implications of the behavioral evidence for fundamental questions like the ones posed above.

Let us begin with the parallel distributed processing framework (Mc Clelland & Rumelhart, 1986 ; Rumelhart & Mc Clelland, 1986). It is a framework for building explicit computational models of human cognitive function. In this framework, we begin by assuming that cognition takes place via the interactions of a large number of simple but highly interconnected computational elements organized into groups or modules (*fig. 1*). Each element is something like a neuron, in that et aggregates inputs that it receives from each other element via excitatory (positive) and inhibitory (negative) connections. The sum of all of these influences in turn determines the activation of the unit, according to a simple non-linear function like the one shown.

In systems of this type, what we might call the active representation of information takes the form of the pattern of activation over the units in the network. So, for example, in the network illustrated in the figure, the presentation of a word would give rise to a pattern of activation over several pools of units, one representing perhaps the alphabetic content of the word, another its semantic content, and a third its phonological content.

We think of the processing of information as the propagation of activation among the units. In the brain, we assume that this is an iterative process, and we simulate this by dividing time into many small steps, and updating the activation of each unit once in each timestep, based on the activations of other units at the previous step. In the case

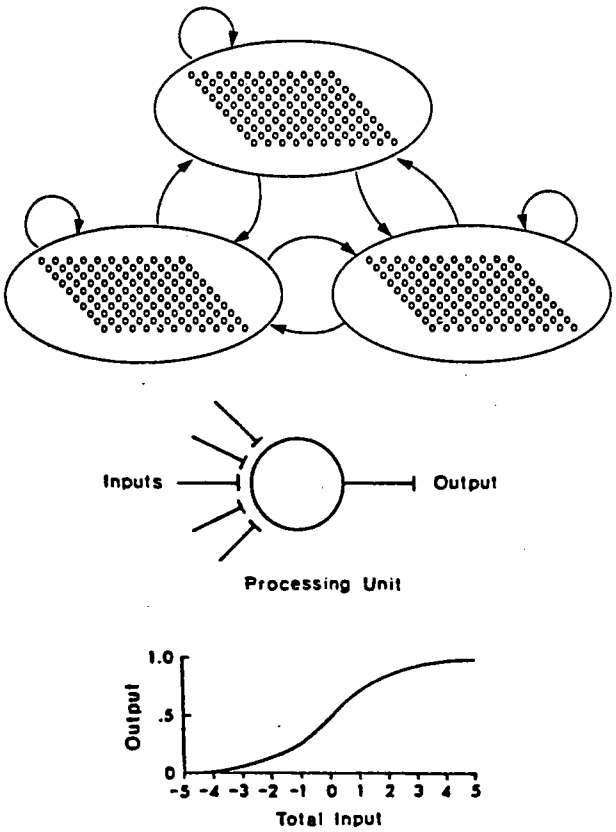


FIG. 1.– The parallel-distributed processing framework.

Le cadre du « parallel-distributed processing ».

Middle and bottom sections after Sejnowski & Rosenberg, 1987.

of networks with symmetrical connectivity, this causes them to settle, after a pattern is presented, into a stable or *attractor*, in which the relevant patterns are active in all parts of the network. In general the framework allows for the possibility that cognitive states are represented as trajectories or patterns of activation that change through time. For the present our interest focuses on the simpler, but still interesting case of attractor networks.

Now, given these conceptions of representation and processing, let us consider knowledge and learning. Where is the knowledge in a parallel distributed processing system ? The answer is, it is stored in the strengths of the excitatory and inhibitory connections among the processing units. So, for example, the knowledge that allows us to translate an alphabetic pattern into a phonological pattern is stored in the strengths of the connections in the pathway leading from the orthographic to the phonological modules. These weights, as we shall see, are capable of storing the complex pattern of regular and exceptional relations between the spellings of English words and their sounds. Other sets of weights can capture the more arbitrary relations between spelling and meaning or between meaning and sound.

Let us briefly consider explicit memories – for example the knowledge that ' Charcot died 100 years ago ' or ' An oak is a tree '. According to the connectionist approach, even such memories as these are stored in connection weights. Of course, when we think explicitly about one of these propositions, we hold it in mind as a pattern of activation. But the knowledge that allows us to reinstate this proposition from memory when appropriately cued is to be found, we believe, in connection weights. We will examine a model that encodes propositional knowledge in this way in a moment. Also, when we think about specific episodic memories, we reinstate a pattern of activation representing visual aspects of the event and its context and of the feelings and emotions of ourselves and others. Again such memories are patterns of activation when we recall them but the knowledge that allows this recall is, we claim, stored in connection weights.

If, as we have suggested, procedural, semantic, and episodic knowledge are all stored in connection weights, then clearly, all three types of learning occur through connection strength adjustment. This fact links the psychological study of learning and memory with the physiological investigation of synaptic modification. For our purposes, we will not consider in great detail exactly what these physiological mechanisms are. Rather, we will rely on the use of a procedure called back propagation that allows us to adjust the strengths of the connections among simple processing units according to a very simple principle called error correction. Basically, this principle says :

Adjust the strength of each connection in proportion to the extent that its adjustment will reduce the discrepancy between the response of the network and external teaching signals.

For example, we can imagine a child learning to translate from spelling to sound, seing a word and then hearing its correct pronunciation. We would use the pattern produced by the visual input to produce a pattern of activation over the alphabetic units : then we would use the existing connection weights to produce a pattern corresponding to the networks representation of the sound. We would compare this to the sound supplied by the teacher, and then adjust each connection weight so as to reduce the difference between the pattern produced by the network and the pattern specified by the teacher.

We now have a framework for thinking about representation and processing ; and about knowledge and knowledge acquisition. I will now briefly describe the results of two simulations demonstrating the power of this approach for the acquisition of both procedural and semantic knowledge.

First we consider procedural knowledge. The network shown in *fig. 2* was developed by Plaut, McClelland, and Seidenberg (1992 ; Plaut & McClelland, 1993) to learn to translate alphabetic patterns representing the spellings of words into phonological patterns representing their sounds. The network consists of input, hidden, and output units ; as well as connections as illustrated in the figure. The training corpus consisted of a set of 3000 monosyllabic

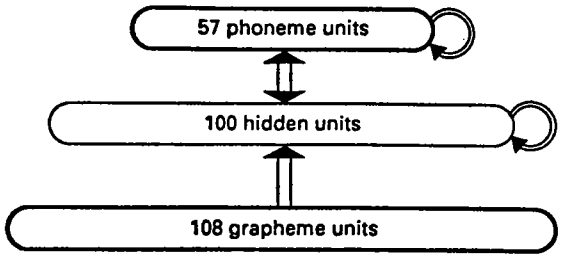


FIG. 2.– The network used by Plaut and Mc Clelland (1993).

Réseau utilisé par Plaut et Mc Clelland (1993).

words. Training proceeded very gradually, capturing the gradual nature of the acquistion of reading skill ; in fact the network was exposed to the entire corpus of training examples 3200 times, with presentations of more frequent words given greater weight than presentations of less frequent words. After training, the network could read 99.7 p. 100 of the words in the corpus correctly, missing only 10 words that were low in frequency and very exceptional either in their spelling or in their pronunciation given the spelling – two examples are the French words ' sioux ' and ' bas ', which are very rare and completely inconsistent with English pronunciation. For the vast majority of words, including most exceptions like those shown on (*fig. 3*), the network was correct. The network was also tested with many pronounceable nonwords, and scored as well as normals in providing plausible pronunciations – as high as 98 p. 100 correct on some sets of items. Based on comparisons of the networks performance with typical adult English speakers, we can say that the network has gradually learned both the regular mapping between spelling and sound as well as the commonly occurring exceptions to it in a way that corresponds to human capabilities in this regard. It has mastered, in short, the procedural skill of translating from spelling to sound – through gradual, incremental learning.

Let us now consider a model that learns semantic information. This model was constructed by Rumelhart (1990) to demonstrate how structured semantic representations can arise in PDP systems from experience with propositions about objects and their properties. The network was trained to capture the information standardly stored in the

Example exception words mastered by the network of Plaut and McClelland (1993) :

HAVE	ONE
PINT	AISLE

FIG. 3.– Some examples of exception words learned by the network of Plaut & Mc Clelland (1993).

Quelques exemples de mots d'exception appris par le réseau de Plaut et Mc Clelland (1993).

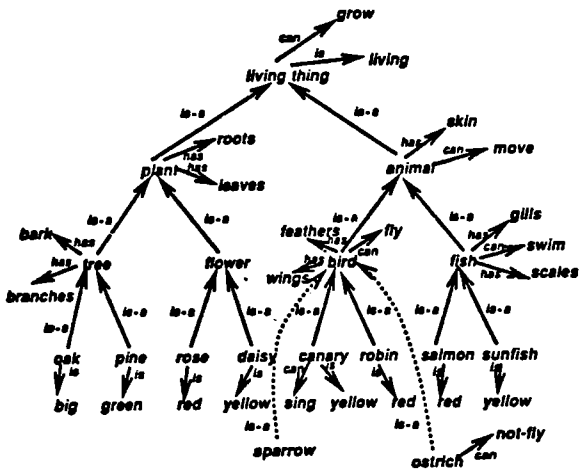


FIG. 4.— A semantic network of the type used in symbolic models of cognition. From Rumelhart and Todd, 1993.

Réseau sémantique du type utilisé dans les modèles symboliques de cognition. D'après Rumelhart et Todd (1993).

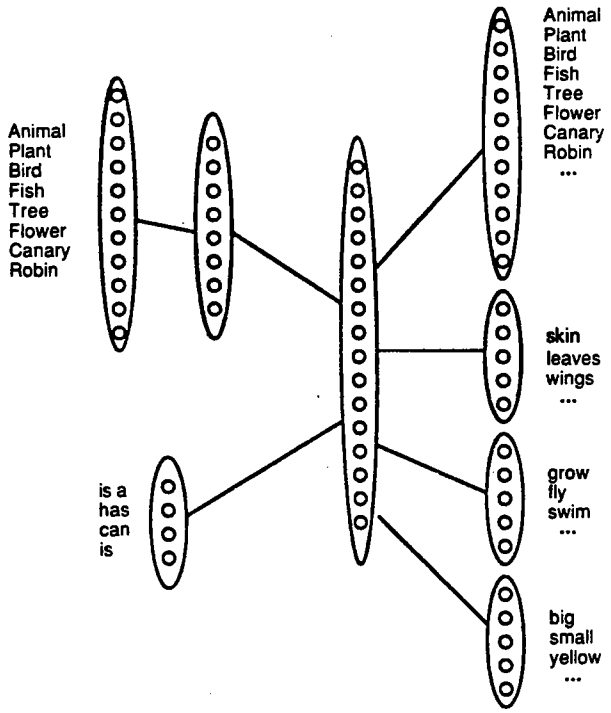


FIG. 5.— The connectionist network used by Rumelhart to learn the facts stored in the previous figure.

Le réseau « connexioniste » utilisé par Rumelhart pour apprendre les faits stockés dans la figure précédente.

old-fashioned semantic networks of the kind illustrated in the *fig. 4.* The goal is to allow the network to store information about concepts in such a way as to permit to generalize from what it has learned about some concepts to other related concepts. In the old approach, this was done by storing information true of a class or subclass of concepts at the highest possible level of the tree. In the new approach, this is done by gradually learning to assign each concept a distributed representation, capturing its similarity relations to other concepts.

Rumelhart's network is shown in *fig. 5.* It consists of a set of input units, one for each concept and one for each of several relations: « IS A », « HAS », « IS », and « CAN ». At the output it consists of a set of units for various completions of simple propositions, such as « ROBIN IS A BIRD », « ROBIN CAN FLY », etc. In between there are two layers of hidden units, one to represent the semantic relations of the concepts and another to combine them with the relations to activate the correct outputs.

Rumelhart trained this network with a set of propositions involving the concepts shown in the previous figure. The training was one input pattern for concept-relation pair, and the task was to turn on all the output units representing correct completions of the proposition. For example, when the input is ' ROBIN IS A ' the correct output is BIRD, ANIMAL, LIVING THING. The entire set of patterns was presented to the network many times, making small adjustments to the strengths of the connections after the presentation of each pattern.

After the network has mastered the training set, it is possible to examine the representations that the network has learned to assign to the words in the concepts. We can look either at the patterns themselves, as well as their similarity relations. I have repeated the simulation myself, and the results that I obtained are presented in *fig. 6.* The figure demonstrates that the network has assigned similar patterns to similar concepts — so for example, the pattern for oak is similar to the pattern for pine, the pattern for daisy is similar to the pattern for rose, etc.

Now we may consider how we may achieve generalization in this network. We may do so if we can assign to a new pattern a representation that is similar to the representation of other similar concepts. To illustrate this, Rumelhart trained the network to produce the correct class inclusion output for the concept sparrow. He simply presented ' Sparrow is a ' as input, trained the network to produce the correct response = ' Bird, Animal, Living Thing '. This caused the network to assign a pattern to sparrow very close to the patterns for robin and canary. As a result the network was able to answer reasonably when probed with other propositions involving sparrows. It said the sparrow can grow and can fly, it has wings, features and skin, and it is small. It was unsure about whether the sparrow could sing and about its color. Thus the network learned to inherit knowledge from what it had learned about other birds without direct instruction.

The point here is that PDP models, trained slowly via interleaved presentation on a representative sample of an entire domain of knowledge, can gradually acquire kno-

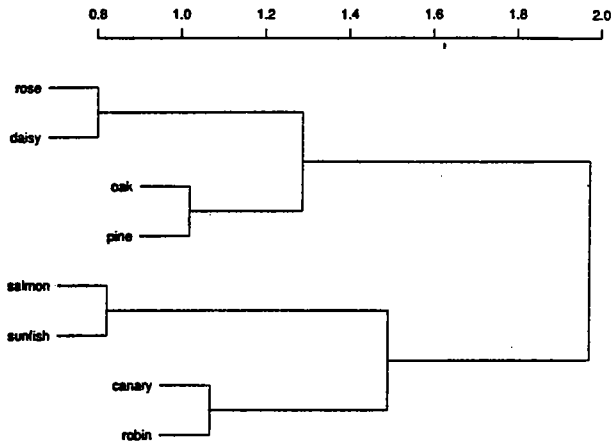


FIG. 6.– Results of a cluster analysis on the representations of each of the specific concepts learned by Rumelhart's network (replication from Mc Clelland, Mc Naughton, and O'Reilly, 1993).

Résultats d'une analyse en grappe portant sur les représentations de chacun des concepts spécifiques appris par le réseau de Rumelhart (d'après Mc Clelland, Mc Naughton et O'Reilly, 1993).

wledge of structured procedures such as spelling-sound correspondence and structured semantic domains such as the domain of living things.

In summary, we have seen how PDP models provide a way not only of representing procedural and semantic knowledge but of indicating how this knowledge may be acquired through gradual experience. In the next section of the talk, I will show how parallel distributed processing models can help us understand two very striking neuropsychological phenomena that have implications for questions about the nature of the organization of human memory.

First we consider a question that arises within semantic memory: How is semantic memory organized in the brain? Is it organized by category, for example with representations of living things stored in one part of the brain and knowledge of man-made artifacts in another? Or is it organized perhaps by modality, with representations of visual properties stored in one part of the brain and knowledge about other types of properties, such as function and use in other parts?

This question has arisen within the domain of neurology due to some fascinating observations reported in a series of studies by Warrington and her colleagues (Warrington & Mc Carthy, 1983, 1987; Warrington and Shallice, 1984). Warrington and Shallice (1984) reported two patients who showed a dramatic impairment in semantic knowledge of living things coupled with only slight deficits in knowledge of man-made objects such as a briefcase or an umbrella. In one test, patients were asked to define concepts when given the word for the concept *Fig. 7* shows the data, along with some example responses.

Clearly, the patients retain more information about man-made objects than they retain about living things. The same results are obtained when the input is a picture so the

Performance of Two Patients With Impaired Knowledge of Living Things on Various Semantic Memory Tasks

Case	*Living thing*	*Nonliving thing*
	Picture identification	
JBR	6 %	90 %
SBY	0 %	75 %
	Spoken word definition	
JBR	8 %	79 %
SBY	0 %	52 %

JBR	Parrot : dont' know	Tent : temporary outhouse, living home
	Daffodil : plant	Briefcase : small case used by students to carry papers
	Snail : an insect animal	Compasse : tools for tellling direction you are going
	Eel : not well	Torch : hand-held light
	Ostrich : unusual	Dustbin : bin for putting rubbish in
SBY	Duck : an animal	Wheelbarrow : object used be people to take material about
	Wasp : bird that flies	Towel : material used to dry people
	Crocus : rubbish material	Pram : used to carry people, with wheels and a thing to sit on
	Holly : what you drink	Submarine : ship that goes underneath the sea
	Spider a person looking for things, he was a spider for his nation or country	Umbrella : object used to protect you from wather that comes

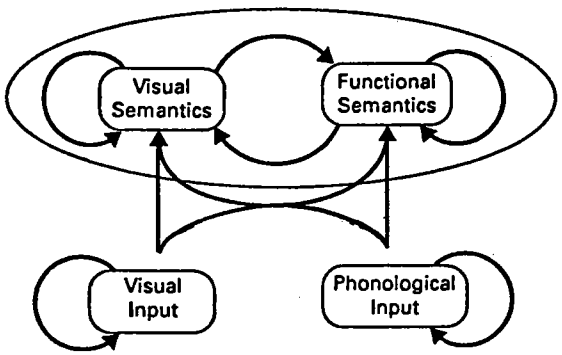
FIG. 7.– Performance of two patients with impaired knowledge of living things in various semantic memory tasks. From Farah & Mc Clelland (1991) based on data in Warrington & Shallice (1984).

Performance de deux malades ayant une mauvaise connaissance des êtres vivants dans différentes tâches de mémoire sémantique. D'après Farah & Mc Clelland (1991), fondé sur les données de Warrington & Shallice (1984).

problem appears to be semantic, not simply one of accessing meaning from sound. Other patients, tested by Warrington and Mc Carthy (1983), show a reverse pattern, although it is not as dramatic.

Taken together the results suggest that localized lesions to the brain can selectively interfere with knowledge of living things on the one hand and with knowledge of man-made objects on the other. It is tempting to conclude from this in fact knowledge of living things is stored in a different part of the cognitive system than knowledge of man-made objects. In fact, though, Warrington and her colleagues (Warrington & Mc Carthy, 1983; Warrington and Shallice, 1984) proposed at first a very different interpretation. On this view, these selective deficits reflect organization by type of information rather than by the type of object. These authors noted that living things seem to be distinguished predominantly by their sensory properties, while man-made objects such as umbrellas are distinguished primarily by their function. They suggested that perhaps to access a concept is to access its distinguishing features. If the living things have more visual distinguishing features, then access to concepts about living things would

be impaired if the portion of the brain containing visual feature information was damaged. This view makes a great deal of sense for a number of reasons. We know that the brain is organized in part by modality ; furthermore, in later research it appeared that the dissociation could best be described, not in terms of living things vs. man-made artifacts, but in terms of concepts that are distinguished by their appearance (including jewelry, for example, which is man-made) vs. concepts that are distinguished by the fact that they refer to objects that are easy to manipulate. Yet Warrington later abandoned the idea, because of what seemed to many people to be a devastating critique. The critique rests on the observation that Warrington and Shallice's patients were impaired, not only for questions about the visual properties of living things, but also for questions about the functional properties of living things. If visual and functional properties are stored separately (perhaps in modality specific stores related to perception and action), it is difficult to see why a lesion that affects visual knowledge more than functional knowledge would affect functional knowledge about living things more than functional knowledge about man-made artifacts. Therefore the theory that semantic memory is organized by type of information rather than by category of concepts was abandoned. Martha Farah and I have used a connectionist model to reinvestigate this issue. In particular, we examined whether a PDP model that adheres to Warrington and Shallice's assumptions could simulate the basic findings from the patient data. In our simulation, the representation of an object is a distributed pattern of activation over several modules (*fig. 8*). We have a visual input module for representing the appearance of the object and another module for representing the name. We imagine there are also semantic modules, perhaps several for different types of information. In our simulation we include two semantic modules, one for visual aspects of the representation of a concept and one for functional information about the purpose and use of the object. The model contains reciprocal connections between both input modules to both semantic modules, allowing bi-directional associations between both types of input and both aspects of meaning. There are also recurrent connections within the visual and phonological input representations, allowing each pattern to function as an attractor. Finally, we also see the concepts as consisting of semantic attractors. These involve connections both within and between the different semantic modules.



The Semantic Memory network of Farah and Mc Clelland (1991)

FIG. 8.– The network model used by Farah and Mc Clelland (1991).

Modèle de réseau utilisé par Farah et Mc Clelland (1991).

Before designing representations of living things and man-made objects, we first tested Warrington and Shallice's claim that living things are distinguished primarily by their visual properties and artifacts are distinguished primarily by their functions. To do this, we presented students with dictionary definitions of the living things and artifacts used in their experiments. One group of students underlined words referring to visual properties, and another underlined words referring to function. The results of this were a little different than Warringhton and Shallice had claimed, but very similar : There were many more visual than functional aspects of living things, but about 7 : 1. For artifacts, there were about equal numbers of visual and functional aspects. We designed our representations accrodingly. Each concept was represented with a random feature pattern, but for living things, 7 : 8 of the features were visual, only 1/8 were functional ; for artifacts, 50 p. 100 of the features were visual and 50 p. 100 were functional. In this way we created 10 ' living thing ' concepts and 10 random ' artifact ' concepts. For each concept, the visual and phonological input representations were simply random patterns. For each concept, the model was trained to use either the name or the picture representation as input, and from this to complete the rest of the pattern – this includes both visual and functional semantics, as well as the other input representation. Note that there are no direct connections between input representations, so in order to perform the association between visual and phonological patterns the network had to make use of the connections via the semantic units.

After training, the network is able to complete all of the concepts accurately, both producing the correct semantic representation and reproducing the correct phonological pattern when the picture is presented and vice-versa. Our question, though, is whether the organization embedded in this network is compatible with the neuropsychological evidence of selective deficits of living things after some lesions, but on artifacts after others. There are no modules corresponding to these categories in the model, yet in fact it does allow us to capture the selective deficits seen in patient behavior, by assuming that their lesions selectively affect visual or functional semantics. To demonstrate this, we carried out a series of lesion experiments on the model. We simulated brain damage of different degrees by destroying different random fractions of the semantic units – either visual semantic units or functional semantic units – and then testing the network to see how well it can perform an analog of picture naming. Specifically, we determined whether the network can activate the correct name pattern when given the picture pattern as input. The results for the

effects of lesions to visual semantics are shown in the *fig. 9*. Here, in the first panel, we see a gradual degradation of function that is much greater for living things that for artifacts. There is very little effect on artifacts unless the lesion is quite severe. The magnitude of the effect is quite consistent with the patient data. Now consider the effects of lesions to functional semantics (*fig. 9* second panel). Here we see that there is very little at all on living things, but a moderate effect on artifacts. The model largely reproduces, then, the double dissociations found in the patients. It does so, essentially, by implementing the Warrington and Shallice account of the data. To activate the correct name, one must activate the distinctive semantics of each concept. But for living things, the bulk of the semantic information is visual, and so by destroying the visual semantic units we prevent the activation of enough of the semantics to produce the correct response. For artifacts, the semantic information is distributed about evenly over the two modules so a lesion to either module produces about the same size effect.

We may now consider the objection that caused Warrington to abandon her model. The argument, was that a lesion that affected visual semantics should leave knowledge of the non-visual aspects of concepts unaffected, yet patients who showed a deficit for living things showed a deficit for functional properties of living things, not just their visual properties. But in our model, we see in fact just this very kind of deficit. As the third panel of the figure shows, we see an increasing deficit in the activation of functional semantic representations as we increase damage to visual semantics, and the deficit is more dramatic for living things than for artifacts. The reason is this. Although the semantic features are segregated as two types the knowledge that allows for their activation is distributed throughout the network, and includes knowledge in connections between visual and functional semantics. Thus when the visual semantic units are destroyed, the network is less able to activate the correct functional semantics. The effect is present for both types of concepts, but much greater for living things than for artifacts. Of course, the deficit is not as great as the deficit in visual knowledge of living things. This is in fact exactly the pattern seen in patients. We show the data from one patient tested by Farah *et al.* (1989). The patient showed relatively subtle deficits compared to the original patients. But the largest deficit was in visual knowledge of living things, consistent with a lesion in visual semantics. The patient also showed smaller deficits in visual knowledge of artifacts and in functional knowledge of living things. There was no significant deficit in functional knowledge of living things. This is just the sort of pattern that the model will show after partial damage to visual semantics.

In summary, Farah and I have been able to show with this model that some of the key data relevant to the organization of semantic memory in the brain is entirely consistent with the idea that the brain is organized by the type of features rather than by the type of concept. On this view concepts are in fact highly distributed, and crucially, the activation of a part of the concept depends on the ability to activate other parts. In this way we are able to see how apparent category specificity of knowledge may emerge from a network that is organized by modality or feature type.

I turn now to a very different kind of neuropsychological dissociation, namely the pattern of deficits and spared capacities seen in hippocampal amnesia. My work on this problem with Mc Naughton and O'Reilly is currently in progress (Mc Clelland, Mc Naughton & O'Reilly, 1993). I am sure everyone is aware of the dramatic pattern of deficits seen in patient HM – a pattern that is now apparent in the data of a large number of other patients with selective but extensive lesions to the hippocampus and related

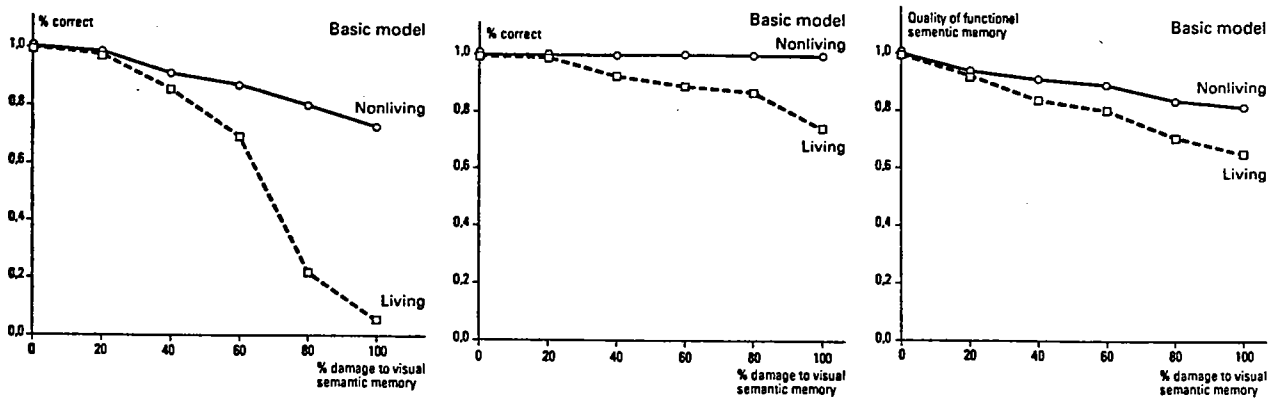


FIG. 9.– Simulation of dissociations in semantic memory from Farah & Mc Clelland (1991). (a) Effects of lesions to visual semantics on naming living things and artifacts (b) Effects of lesions to functional semantics on the same. (c) Effects of lesions to visual semantics on activation of functional semantics.

Simulation de dissociation en mémoire sémantique, d'après Farah et Mc Clelland (1991). (a) Effets de lésions sur la sémantique visuelle concernant les noms d'êtres vivants et d'artéfacts : (b) Effets de lésions sur la sémantique fonctionnelle concernant le même sujet ; (c) Effets de lésions sur la sémantique visuelle concernant l'activation de la sémantique fonctionnelle.

structures in the medial temporal lobes. These patients show a dramatic deficit in the ability to acquire new explicit memories of the contents of specific episodes and events. However, at the same time they are completely normal in the use of their existing semantic and procedural knowledge, and in fact their acquisition of new skills appears to be completely intact. They also show normal repetition priming effects in a wide range of tasks. For example, they show a normal amount of perceptual facilitation in the identification of visually presented words due to a prior presentation of the word. They also show a severe temporally graded retrograde amnesia for episodic information. *Figure 10* shows data on this from the laboratory of Squire (Mac Kinnon and Squire, 1989 ; Squire *et al.* 1989). Here we see a temporally graded retrograde amnesia that appears to encompass 10 years or more. Memories of episodes from childhood and adolescence appear to be independent of the hippocampus, but more recent memories do appear to depend on the hippocampal formation.

Our view of the organization of mammalian memory that is consistent with these facts is illustrated in *fig. 11*. First, we imagine that ultimately all kinds of knowledge can ultimately be stored in the connections among neurons in the cortex and other non-hippocampal structures such as the basal ganglia (for brevity we will refer to this as the cortical system). We imagine that this cortical system learns gradually, with each experience producing very small changes to connection weights. These small changes, we assume, are the basis of repetition priming effects, and they gradually give rise to cognitive skills, such as the ability to translate from spelling to sound, and semantic knowledge, such as that exhibited by Rumelhart's net. However, in addition to this cortical system, we assume that there is also rapid storage of traces of specific episodes within the hippocampus. On this view, when an event or experience produces a pattern of activation in the neocortex, it produces as the same time a pattern of activation at the input to the hippocampus. So for example, if I am introduced to someone, I form a pattern of activation including the appearance of the person, the name, and other aspects of the situation. Synaptic modifications within the hippocampus itself then auto-associate the parts of the pattern. Later, when a retrieval cue is presented (let us say the person reappears and I wish to recall his name), this then produces a partial reinstatement on the hippocampal input pattern. This is then completed with the aid of the modified synapses in the hippocampus, and then reinstated in the neocortex via return projections. The assumption is that these changes are large enough so that with one or a few trials they may sustain accurate recall, while the changes that take place in the cortex are initially too small for that. Gradually, though, through repeated reinstatement of the same trace, the cortex may receive enough trials with the same association to learn it in the neocortical connections. This reinstatement can occur, I propose, either through repetition of the association over many different events ; or through repeated reinstatement from the hippocampus. Thus on this view the hippocampus can serve both as the initial cite of storage but also as teacher to the neocortex.

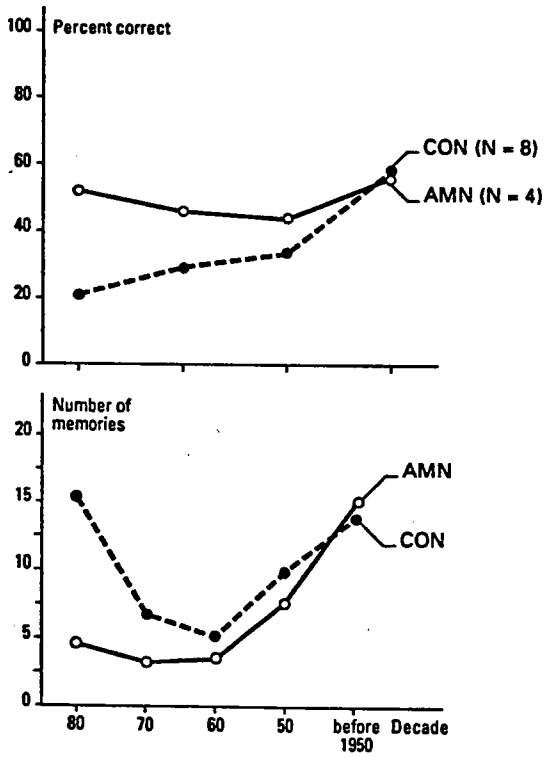


FIG. 10.– Temporally graded retrograde amnesia in humans. Recall performance is shown as a function of the date of the event remembered. (a) Personal episodes (Mac Kinnon and Squire, 1989). (b) Public events (Squire, Haist, and Shimamura, 1989).

Amnésie rétrograde temporellement graduée chez l'homme. Nombre de mémoires rappelées par des sujets normaux et amnésiques. (a) Épisodes personnels (Mac Kinnon et Squire, 1989) (b) Événements publics (Squire, Haist et Shimamura, 1989).

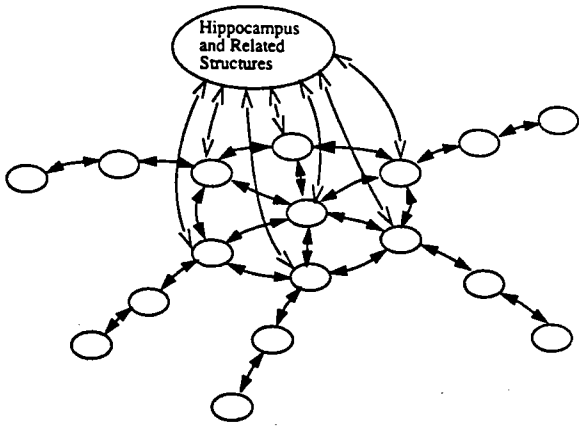


FIG. 11.– Sketch of the human neocortical and hippocampal systems (from Mc Clelland, Mc Naughton, and O'Reilly, 1993).

Schéma des systèmes néocortical et hippocampique chez l'homme (d'après Mc Cleelland, Mc Naughton et O'Reilly, 1993).

The above description provides an account of the data, and although not everyone accepts it, it is not really totally new. Many investigators have suggested some or most of these ideas with varying degrees of completeness. What I do not know of, however, is any very clear story about why the system might be organized in this way. Let me phrase the questions as follows :

First, why is it that we need a special system for rapid storage of episodic memories, if knowledge of all types is eventually stored in connectionist within the neocortex ?

Second, why is consolidation of information into the neocortex so slow ? Why does it require many repeated reinstatements apparently spanning years in some cases ?

The final substantive point of my talk will be to suggest an answer to these questions. The answer requires us to return to the demonstration of the acquisition of procedural and semantic representations that I presented earlier in the lecture. There I argued that the effective discovery of the shared structure that underlies an entire domain such as spelling-to-sound correspondence or the semantic organization of knowledge of living things specifically requires gradual learning, in which the learning of any one association is interleaved with learning about other associations.

Let me illustrate and amplify this point by comparing what happens in Rumelhart's semantic network if we try to teach it some new information in either of two different ways. The first way, which I call focussed learning, involves teaching the network the new information all at once, without interleaving it with ongoing exposure to the structure of the entire domain. The second way, which I call interleaved learning, involves simply introducing the new information into the mix of experiences that characterize the entire domain. As our example, I consider the case of the Penguin. Now as we all know, the penguin is a bird, but it can swim, and it cannot fly. I took Rumelhart's semantic network and taught it these two things in the two different ways just described : The focussed case involved simply presenting the two items repeatedly and watching the network learn. The interleaved case involved adding the two new items to the same training set used previously and continuing training as before. The results of these experiments, for the case of ' penguin can grow-swim-fly ', are shown on the *fig. 12.* Here if we look at the number of presentations required for acquisition, in the left panel of the figure, we see that focused learning is better, since acquisition is considerably faster with focussed learning than it is with interleaved learning. But it turns out that this slight advantage has been purchased at a considerable cost. For it turns out that with focussed learning, the training on the case of the penguin has strongly corrupted the models understanding of other concepts. Indeed if we consider the networks knowledge of the concept robin, in the lower panel, we see that focussed learning has produced a dramatic interference. In fact, the model now thinks that all animals can swim, and even some plants. But in the case of interleaved learning, we see very little interference with what is already known. To be sure there is a slight effect,

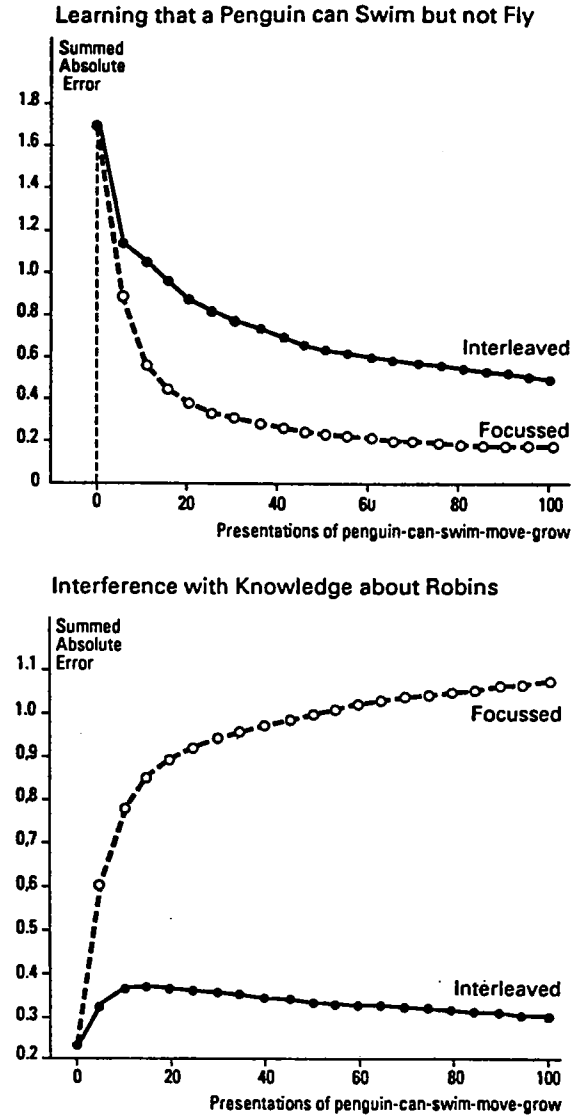


FIG. 12.– Effects of focussed and interleaved learning. (a) Learning that a penguin can swim but not fly. (b) Interference of learning about the penguin with knowledge about the robin (Mc Clelland, Mc Naughton & O'Reilly, 1993).

Effets de l'apprentissage centré et inter-paginé. (a) Connaissance du fait que le pingouin peut nager mais non voler. (b) Interférence de la connaissance concernant le pingouin et de celle concernant le rouge-gorge (Mc Clelland, Mc Naughton et O'Reilly, 1993).

but as interleaved learning proceeds this is even reduced gradually over trials. Thus with focussed learning new knowledge corrupts the structured database in the network ; with interleaved learning new knowledge is gradually added with no disruption.

This demonstration recapitulates a phenomenon that has previously been reported by researchers who have tried to use networks of the type used in Rumelhart's semantic network to model episodic memory (Mc Closkey and

Cohen, 1989). They termed the phenomenon catastrophic interference, and rejected networks of the type that gradually discover structure through incremental learning as models of episodic memory, since storage in episodic memory clearly involves what I have called focussed learning. I agree that back propagation networks are poor models for this kind of learning. However, the kind of learning that they are good at is just the kind of learning we think the cortex is specialized for – gradual discovery of skills from the overall structure of experience.

For the discovery of overall structure, gradual, interleaved learning is necessary, so that the connection weights in the network come to reflect influences of all of the examples in the domain. In fact, it can be shown mathematically that in order to converge on a set of connection weights that best captures the structure of a particular domain, it is necessary in the limit to reduce the learning rate asymptotically to 0; the closer it gets to zero, the better we will approximate correct connection weights. This point leads to the observation that if the brain is to be able to extract the structure, it is going to have to learn slowly.

With these ideas in mind, we can now return to the two questions asked above.

First, why is it that we need a special system for rapid learning of the contents of specific episodes and events, if knowledge of all types is eventually stored in connectionist within the neocortex?

The answer begins with the idea that the cortical system is specialized for the extraction of shared structure of events and experiences. In order to do this, and to avoid catastrophic interference with that structure, it is necessary for the cortical system to use a very small learning rate. In this context, the role of the hippocampus is to provide a system in which the contents of specific episodes and events can be stored rapidly, without at the same time interfering with the structure that has been extracted by the cortical system.

Second, why is consolidation of information into the neocortex so slow? Why does it require many repeated reinstatements apparently spanning years in some cases?

My answer is that consolidation is slow precisely because it would be disruptive if information stored in the hippocampus were incorporated in the neocortex all at once. It is only when new information is gradually interleaved in this way that we are able to incorporate new knowledge into the neocortical system without interfering with what we already know.

Let me conclude by summarizing briefly. First I reviewed the parallel distributed processing framework, and showed how it provided a mechanism for the gradual acquisition of cognitive skills and for the gradual discovery of semantic structure. Then I considered two issues that have arisen within the field of cognitive neuropsychology: First, is semantic memory organized by conceptual category, or by modality of information about each concept stored in memory? Second, why do we have a hippocampus, when ultimately everything that we know is stored in the neocortex? I argued that parallel distributed processing models provide new ways of studying these questions, and indeed I suggested answers to these questions that make sense of the existing neuropsychological facts in terms of mechanisms that are consistent with the principles of parallel distributed processing. In closing I would like simply to say that the particular ideas I have presented are not the main point of this talk. Rather, the main point is to indicate that computational models, based on principles of parallel distributed processing, can help us gain further insight into the neural mechanisms underlying cognition, as these are revealed through the striking dissociations we see in the behavior of neurological patients.

References

Farah M.J., Hammond K.H., Mehta Z. & Ratcliff G. (1989). Category-specificity and modality-specificity in semantic memory. *Neuropsychologia, 27:* 193-200.

Farah M.J. & McClelland J.L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120:* 339-357.

McClelland J.L., Mc Naughton B.L. & O'Reilly R.C. (1993). Why we have complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Manuscript in preparation.*

McClelland J.L., Rumelhart D.E. and the PDP research group. (1986). Parallel distributed processing: *Explorations in the microstructure of cognition. Vol. II.* Cambridge, MA: MIT Press.

Mc Closkey M. & Cohen N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *IN:* G.H. Bower, (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, 24. San Diego: Academic Press.

Mac Kinnon D. & Squire L.R. (1989). Autobiographical memory in amnesia. *Psychobiology, 17:* 247-256.

Plaut D.C. & Mc Clelland J.L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. To appear *in:* Proceedings of the 15th Annual Conference fo the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Plaut D.C., Mc Clelland J.L., & Seidenberg M.S. (1992). Reading words and pseudowords: Are two routes really necessary? Talk presented at the *Annual Meeting of the Psychonomic Society,* St. Louis, MO, November.

Rumelhart D.E. (1990). Brain style computation: Learning and generalization. *In:* S.F. Zornetzer, J.L. Davis, and C. Lau (Eds.). An introduction to neural and electronic networks (pp. 408-420). Academic Press: San Diego, CA.

Rumelhart D.E., Mc Clelland J.L. and the PDP research group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1. Cambridge, MA: MIT Press.

Rumelhart D.E. & Todd P.M. (1993). Learning and connectionist representations. *In:* D.E. Meyer & S. Kornblum, (Eds.), Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp 3-30). Cambridge, MA: MIT Press.

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

Squire L.R., Haist F. & Shimamura A.P. (1989). The neurology of memory: Quantitative assessment of retrograde amnesia in two groups of amnesic patients. *The Journal of Neuro-science, 9 (3):* 828-839.

Warrington E.K. & Mc Carthy R. (1983). Category specific access dysphasia. *Brain, 106:* 859-878.

Warrington E.K. & Mc Carthy R. (1987). Categories of knowledge: Further fractionation and an attempted integration. *Brain, 110:* 1273-1296.

Warrington E.K. & Shallice T. (1984). Category specific semantic impairments. *Brain, 107:* 829-854.