Biol. Cybern. 68: 455–463 (1993)

Biological Cybernetics

A neural model of the dynamic activation of memory

M. Herrmann [1], E. Ruppin [2], M. Usher [3]

[1] Abteilung für Computerwissenschaft der Universität Leipzig, O-7010 Leipzig, Germany
[2] Department of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[3] Computation and Neutral Systems, Division of Biology 216-76, Caltech, Pasadena, CA91125, USA

Received: 14 March 1992/Accepted in revised form: 5 October 1992

Abstract. We study an Attractor Neural Network that stores natural concepts, organized in semantic classes. The concepts are represented by distributed patterns over a space of attributes, and are related by both semantic and episodic associations. While semantic relations are expressed through an hierarchical coding over the attribute space, episodic links are realized via specific synaptic projections. Due to dynamic thresholds expressing neuronal fatigue, the network's behavior is characterized by convergence toward the concept patterns on a short time scale, and by transitions between the various patterns on a longer time scale. In its baseline, undamaged state, the network manifests semantic, episodic, and random transitions, and demonstrates the phenomen of priming. Modeling possible pathological changes, we have found that increasing the 'noise' level or the rate of neuronal fatigue decreases the frequency of semantic transitions. When neurons characterized by large synaptic connectivity are deleted, semantic transitions decay before the episodic ones, in accordance with the findings in patients with Alzheimer's disease.

1 Introduction

The dynamic organization of memory, an essential mechanism underlying thought processes, is composed of two interacting elements: memory structure and retrieval dynamics. These components can be studied at two levels of description, a mental level involving activation of mental states, and a neural level involving their realization in the brain. Recently Gröbler et al. (1991) have proposed a scheme for associative free recall based on a mental architecture, where concepts and the associations between them are represented by nodes in a weighted graph and their links, respectively. In this work we discuss the processes underlying the dynamics of memory retrieval from a perspective based on a neural level of description.

We present a biologically motivated neural network model that is able to generate basic characteristics of normal memory function, such as semantic and episodic associations. Following the growing evidence supporting the diffuse modulatory role of 'lower' brain stem neuronal populations on 'higher' cortical activity (Mamelak and Hobson 1989; Sutton et al. 1992), it is shown that the level of 'noise' manifested in the neurons' dynamics influences the level of transitions actually taking place via the semantic associations. Motivated by neuronanatomical findings in Alzheimer's disease, pathological changes taking place on the neural level are modeled, and are shown to lead to a pattern of memory failure resembling some neuropsychological reports.

In this work we investigate the *declarative* memory system, which is assumed to store factual, non-operational, information about the world surrounding us (Squire 1982; Schacter 1989). Memorized concepts are typically related via two types of connections, *semantic* and *episodic*. While *episodic* associations are formed between concepts that have significant spatiotemporal relations, *semantic* associations reflect abstract relations between the memorized concepts, that are not necessarily acquired as part of the personal history (Tulving 1985). The stored concepts are hierarchically organized into semantic classes; specific concepts at a lower level of the hierarchy, (e.g., cats, dogs), are grouped into classes of more general concepts (e.g., animals).

Several approaches can be found in the literature of semantic memory dealing with concepts organization. Originally, it was suggested that concepts are organized in an hierarchical tree (Collins and Quillian 1969). According to this approach, each concept 'inherits' the attributes that are stored at his ancestral nodes in the tree (in addition to its own specific attributes), satisfying therefore an 'economy principle'. For example, the attribute 'eats' is stored with the 'animal' concept and inherited by the 'cat' and 'dog' concepts. This work follows an alternative approach proposed by Smith et

Correspondence to: E. Ruppin

al. (1974), where the memorized concepts are stored as distributed vectors of attributes. Thus, semantic proximity and attributes inheritance are naturally obtained due to similarity in the encoding over the semantic space. The semantic space is subdivided into aggregates of closely related concepts, denoted as *semantic classes*. Episodic associations, on the other hand, are not related to the semantic similarity between the concepts involved. We hence assume that episodic associations are formed by explicit synaptic projections between the neurons composing their corresponding patterns. While the original, tree-like model of semantic memory could be easily implemented in a 'localistic' connectionist architecture (Fahlman 1981), were each concept is represented by one neuron, the approach of Smith et al. is more suitable for networks with distributed representations. The latter have several computational advantages over the localistic ones (Hinton 1981; Rumelhard and McClelland 1986), and are more plausible biologically; empirical evidence shows that distributed patterns of activity are used for various information processing tasks including memory encoding (Heit et al. 1988; Tanaka and Saito 1991), olfaction (Skarda and Freeman 1987) and motion computation (Georgopolous et al. 1988).

The network presented manifests some dynamical aspects of memory that have been widely discussed in the psychological literature. Empirical studies (Anderson 1985; Collins and Quillian 1969; Ratcliff and McKoon 1981) have shown that semantic proximity can influence the frequency and speed of associations. This is the phenomenon of *priming*, according to which activation contributed by an input (priming) concept facilitates the response time to a target concept if they are semantically related. The magnitude of facilitation was shown to depend on the time delay between the priming and target stimulus (Ratcliff and McKoon 1981). The most popular framework addressing temporal aspects of associations is the *Spreading Activation theory* (Collins and Loftus 1975; Anderson 1976, 1985). According to this theory, activation reflecting mental activity spreads in parallel from all active concepts to the other concepts with which they are related to. However, Spreading Activation cannot solely account for the complexity of human mental processes which manifest both parallel and serial characteristics.

Following Kihlstrom (1987), we assume that mental processes can be divided into *conscious* and *unconscious* processes, and that while unconscious processes are parallel, conscious attentional processes are serial (Treisman 1980). While the classical approach considers the conscious and unconscious aspects of cognition as manifested in different memory stores (Schacter 1989), a more parsimonious approach, according to which the conscious-unconscious dichotomy is a manifestation of a unitary system has been advanced by Gröbler et al. They proposed that conscious states are patterns of neutral activity that are distinguished from unconscious states by a threshold of activity. However, this requires that the spreading activation should be accompanied by an auxiliary mechansim, since otherwise the activation would spread uniformly all over the semantic space. Collins and Loftus (1975) have originally proposed that the 'intersection' of the spreading neural activity is detected and leads to the seriality required for generating a specific response. However, their original formulation has been computer oriented rather than biologically plausible, requiring further processing by some higher level system. Gröbler et al. have proposed a more biologically oriented mechanism, involving different dynamics for sub and supra-threshold activation, and leading to autonomous dyanmics with both parallel and serial characteristics.

In the next section we discuss the framework of Transient Attractor Neural Networks (TANN), that enables us to give a unified and natural model encompassing sequences of parallel and serial concept activation. Our model is presented in Sect. 3, and its 'baseline' dynamic behavior is presented in Sect. 4, together with the priming phenomenon. In Sect. 5, the performance of a damaged network is examined, modeling some neuroanatomical and neuropsychological findings in Alzheimer's disease. Finally, in the last section, the potential of a neural level of description in TANNs is dicussed.

2 Cognitive modeling with TANN

Our model is based on a generalization of an *Attractor Neural Network* (ANN) (Hopfield 1982; Amit 1989). An ANN is an assembly of formal neurons connected recurrently by synapses. The neuron's state is a binary variable S, taking the values ± 1 denoting firing or resting states. The network's state is a vector specifying the binary values of all neurons at a given moment. Each neuron receives inputs from all other neurons to which it is connected, and fires only if the sum of the inputs is above its threshold. This process may include a stochastic component (noise) which is analogous to temperature T in statistical mechanics. When a neuron fires, its output, weighted by the synaptic strength, is communicated to other neurons and as a consequence, the network's state evolves. Using specific learning rules (which govern the synaptic strength), the stored memory patterns are made *attractors* of the network's dynamics, so that the network converges to a memory state if a similar pattern is presented as an input.

ANN models of memory and associations have been previously presented by Hoffman (Hoffman 1987; Hoffman and Dobscha 1989), in an attempt to model thought disorders manifested in various psychiatric diseases. However, these models cannot capture the complexity required for modeling associative thought processes, being based on an ANN in which all patterns are equidistant, and whose dynamics is basically reduced to convergence into attractors. Our work is an attempt to further extend Hoffman's approach, by introducing a complex dynamical system characterized by two time scales, and a hierarchical metric structure of memory concepts.

Using a Transient ANN (TANN), a richer dynamical behavior is achieved. On a short time scale, a TANN (like an ANN) converges toward an attractor. However, transitions between the various attractors (therefore denoted *transient attractors*) take place, albeit on a longer time scale (Horn and Usher 1989, 1990). Such dynamical systems have been recently proposed in the Neutral Network literature on the basis of synaptic delays (Kleinfeld 1986; Sompolinsky and Kanter 1986), neural adaptation (Horn and Usher, 1989, 1990), or slow inhibition (Abbott 1990). The dynamics of an TANN (as well as of ANN) are governed by a nonlinear mechanism of competition among the concept patterns, and thus while many concept patterns are partially activated concomittantly, reflecting a parallel unconscious mental activity, the competition and the convergence to an attractor state inherently leads to the seriality of conscious states. Since our model is based on a distributed representation, the spreading of activation takes place in the concepts' and not in the neurons' space. The network's dynamics inherently yield a winner-take-all mechanism ensuring seriality, and therefore a consciousness threshold mechanism less arbitrary than the one proposed previously in Gröbler et al. (1991), is obtained.

Several schemes for storing hierarchical patterns in ANN have been previously proposed, distinguished via the way patterns are encoded, and their rule for synaptic storage. Feigelman and Ioffe (1987) and Gutfreund (1988) have proposed storing schemes for hierarchical patterns in ANN which eliminate the correlation between patterns from the same class. This method, while enhancing the patterns' stability and increasing the network's capacity, removes all bias towards intraclass associations, and thus is unable to reflect semantic proximity. Another storing scheme for hierarchical patterns which eliminates global correlations, but not intraclass ones, was proposed by Tsodyks (1990) and Herrmann and Tsodyks (1991). According to this scheme, patterns are encoded in such away that concepts in one class have a common core of defining properties. However, while in these models the common core is fixed, experimental results on concepts' similarity show that there are almost no attributes common to all concepts in one class (Rosh and Mervis 1975), and hence classes of concepts should be characterized by *fuzzy cores*. Another scheme for encoding classes of concepts has been proposed by Ritter and Kohonen (1989), whereas semantic relations are reflected as metric distances on a 2-dim surface. Our model is an extension of that proposed by Herrmann and Tsodyks (1991); concepts belonging to the same semantic class have a fuzzy common core, and both the semantic and episodic associations are examined.

3 The model

Consider a newtork of N neurons, characterized by two-valued variables $S_i \in \{0, 1\}$ corresponding to a non-active or an active state. Each neuron is subject to a dynamical threshold variable θ_i (Horn and Usher 1989, 1990). According to our scheme, the neurons correspond to properties or attributes, and the distributed patterns of neural activity stand for concepts. Each semantic class is represented by a vector $\xi^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, where $\xi_i^\mu = 1$ if at least one of the concepts in the μ-th class has the i-th property, and $\xi_i^\mu = 0$ otherwise. We first construct the 'class' vectors ξ^μ by

$$\xi_i^\mu = \begin{cases} 1 & \text{with probability } p_1 \\ 0 & \text{with probability } (1-p_1)\,. \end{cases} \tag{1}$$

In each class we then generate several concepts; the vector $\xi^{\mu\nu}$ that stands for the ν concept in class μ is generated by

$$\xi_i^{\mu\nu} = \xi_i^\mu \eta_i^{\mu\nu}, \qquad \text{where} \tag{2}$$

$$\eta i^{\mu\nu} = \begin{cases} 1 & \text{with probability } p_2 \\ 0 & \text{with probability } (1-p_2)\,. \end{cases} \tag{3}$$

The probabilities p_1 and p_2, and L_1 (number of classes, L_2 (number of individual concepts per class), are parameters of the model. The probabilities p_1 and p_2 are chosen to be smaller than half, reflecting the biological constraint that neural representations are probably sparse. This guarantees that concepts in different classes will have very few attributes in common. The obtained patterns have the average activity $\langle \xi^{\mu\nu} \rangle = p_1 p_2 \equiv p$. Patterns in the same semantic class have some common attributes, however there is a very low probability ($p_2^{L_2}$), that a specific attribute will be shared by all the concepts in the class. The strengths of the synaptic connections are defined as in Tsodyks (1990):

$$J_{ij} = \frac{1}{(1-p)pN} \sum_{\mu=1}^{L_1} \sum_{\nu=1}^{L_2} (\xi_i^{\mu\nu} - p)(\xi_j^{\mu\nu} - p) \tag{4}$$

The dynamic equations for the variables S_i are given via the post synaptic potentials h_i

$$h_i(t+1) = \sum_{j=1}^{N} j_{ij} S_j(t) - \lambda(p - M(t)) \tag{5}$$

$$S_i(t+1) = \begin{cases} 1 & \text{with probability } 1/ \\ & (1 + \exp(-(h_i(t+1) - \theta^0 - \theta_i(t))/T)) \\ 0 & \text{with probability } 1/ \\ & (1 + \exp(h_i(t+1) - \theta^0 - \theta_i(t))/T)) \end{cases} \tag{6}$$

where T is the thermal noise, level θ^0 denotes a constant threshold and λ is a confinement parameter which regulates the network overall activity $M = \langle S_i \rangle$ to its mean value p.

The equation for threshold dynamics is (Horn and User 1989, 1990):

$$\theta_i(t+1) = \theta_i(t)/c + bS_i(t+1) \tag{7}$$

According to this equation, while a neuron is active, its dynamic threshold $\theta_i(t)$ increases asymptotically to the value $\theta_{\max} = cb/(c-1)$ and deactivates the corresponding neuron, thus expressing neuronal fatigue. The

parameters b and c ($c > 1$) represents the rate of increase and decay of the dynamic thresholds, respectively. The role of the dynamic thresholds is to provide a mechanism of motion in the concept space; neurons that are active for a relatively long time are deactivated temporarily, and the network's state evolves into a new pattern.

We define the normalized retrieval qualities of the patterns as

$$m^{\mu\nu}(t) = \frac{1}{p(1-p)N} \sum_{i=1}^{N} (\xi_i^{\mu\nu} - p) S_i(t)$$

The latter are macroscopic thermodynamic variables that are correlates of concepts' activation in the network. A concept will be considered activated when its corresponding retrieval quality exceeds some *consciousness threshold* (chosen to be 0.9), whereas the retrieval qualities of all other patterns are below an *unconsciousness threshold* (chosen as 0.5). One should notice however, that unlike in Gröbler et al. (1991), these thresholds play no active role in the dynamics.

2 Semantic transitions and priming

There are three kinds of transitions that can occur in the network; *semantic transitions*, occurring between concepts belonging to the same semantic class, *episodic transitions* taking place between concepts belonging to distinct semantic classes by are linked together by episodic associations, and *random transitions* occurring between concepts that are not related either schematically or episodically.

We first study the properties of our model restricting ourselves to semantic transitions. The dynamic behavior of the model is illustrated by simulations of a network which stores three classes of three concepts, i.e. $L_1 = L_2 = 3$. A characteristic example in terms of the retrieval qualities is shown in Fig. 1. Each stripe presents the retrieval qualities of concepts of a different semantic class. Due to the competition among the patterns, most of the time the network's activity is dominated by one of the patterns, leading to a seriality effect. However, due to the threshold adaptation, the activity of the dominant pattern decays and the network's state converges to another pattern. Semantic transitions are more frequent than the random transitions occurring across classes, leading to *semantic bias* that is characteristic of human memory function.

The extent of the semantic bias is influenced by the thermal noise T. For low T almost no transitions occur at all, whereas strong noise destroys the retrieval capabilities of the network. We have run simulations of the network behavior in which we have varied the temperature T. As shown in Fig. 2, the average fraction of intraclass transitions turned to be higher for low noise. For high T values, the fraction of semantic transitions approaches the baseline of randomness given by $(L_2 - 1)/(L_1 L_2 - 1) = 1/4$. In a similar fashion, the magnitude of semantic bias decreases as the value of the asymptotic $\theta_{\max}$ is increased.

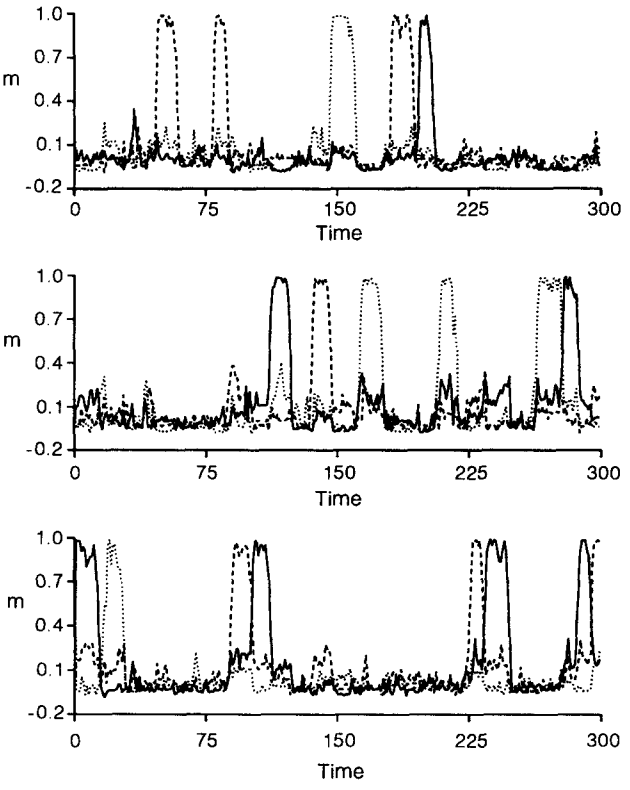

Fig. 1. Illustration of the dynamic behavior of the model. Each strip shows the retrieval qualities of concepts in a semantic class, as function of time. Semantic bias is evident. Parameters: $p_1 = 1/3$, $p_2 = 1/4$, $\theta_{\max} = 0.4$, $\theta_0 = 0.35$, $c = 1.2$, $\lambda = 0.75$, $N = 500$

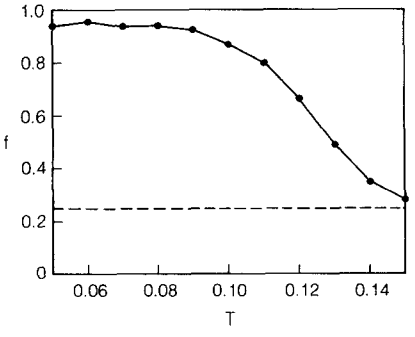

Fig. 2. Averaged fraction of semantic transitions as a function of thermal noise strength

Our investigation of the influence of temperature level on the network's behavior have been motivated by the findings supporting the existence of aminergic modulation of cortical activity (summarized in Mamelak and Hobson 1989; Servan-Schreiber et al. 1990; Sutton et al. 1992). Interpreting this modulation as a variation in the level of the noise in the network, Mamelak and Hobson (1989) claim that neuromodulatory changes leading to increased neuronal noise level may lead to the bizarre dreams appearing in REM sleep, characterized by incongruent 'leaps' from one theme to another. In accordance with their proposal, we show that as the

noise level is increased, the semantically related associations are gradually replaced by random transitions. Hoffman (1987) has previously claimed that an increase in the noise level would probably not result in severe thought discontinuities denoted as 'loosening of associations', but in a milder form of though disturbances known as 'flight of ideas'. As shown here, in a more intricate model than the basic ANN Hopfield model used by Hoffman, variations in the noise level may indeed undermine the fraction of congruent, semantic transitions.

Priming is investigated in our framework by defining the response time as the number of the time steps elapsing from the moment an input stimulus is applied to the network, until this stimulus dominates the network's activity. Applying a stimulus to the network is modeled by adding a new term, proportional to one of the memory patterns, to the postsynaptic potential,

$$h_i^{\text{prim}}(t_0) = h_i(t_0) + \xi_i^{\mu\nu} \tag{8}$$

where the 'stimulus strength' $\varepsilon \ll 1$ accounts for sensory coupling.

In the first simulation experiment the network was initialized by a pattern $\xi^{\mu\nu}$. Thereafter, a sensory input field (Eq. 8) proportional to $\xi^{\mu'\nu'}$ (with $\nu \neq \nu'$) is applied. We measured the average response time as function of ε, for $\mu' = \mu$ (same semantic class), and for $\mu' \neq \mu$ (control). In Fig. 3, the ratio of the respective values of response times for the priming vs. control stimuli, as a function of ε, is shown. For very small values of ε, the sensory input is negligible, while for higher values ($\varepsilon \to 1$) the transition is almost instantaneous in both cases. At intermediate ε values semantic transitions are significantly more rapid than the controls. As shown in Fig. 3, this trend is further emphasized as the patterns' encoding is more sparse (p_2 decreases).

In a second simulation, we have sequentially applied two input patterns belonging to the same semantic class. The first input represents the priming stimulus, and the second one the target.

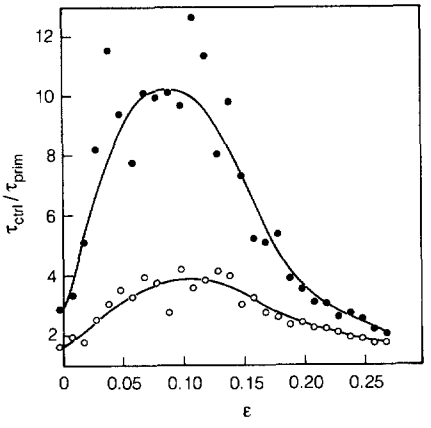

Fig. 3. Priming. The ratio τ_{ctrl}/τ_{prim} is given as a function of the strength of sensory coupling for different activities (lower curve: $p_2 = 1/4$; upper curve: $p_2 = 1/6$) τ_{prim} and τ_{ctrl} are the averaged response times in the priming and control experiment, respectively. The curves are obtained by performing statistical regression analysis of the simulation results

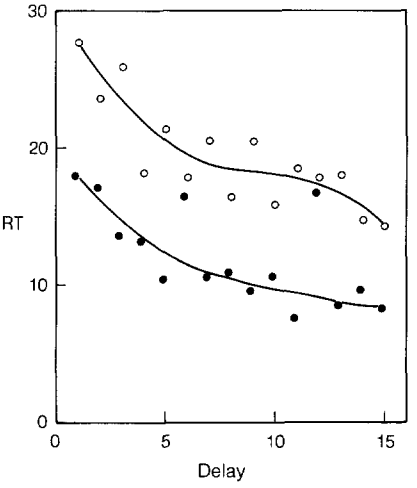

Fig. 4. Priming. Averaged retrieval time of a target pattern as a functions of the delay between the priming and the target pattern. (upper curve: $\varepsilon = 0.075$; lower curve: $\varepsilon = 0.125$) The curves are obtained by performing statistical regression analysis of the simulation results

We have measured the convergence time to the second pattern, as function of the delay time betweeen the two inputs. As can be seen in Fig. 4, the speedup in the convergence time builds gradually with the delay between the two patterns, as in Ratcliff and McKoon (1981) experiments. However, according to our model, the explanation of this effect is not a 'gradual accumulation of activation' at the second concept, but is done to the seriality and the bias in transition probabilities inside semantic classes. The first input pattern tends to activate its corresponding concept. However, if the delay time between the two inputs is smaller than the characteristic time for concepts activation, the impact of the second input on the network is diminished. The response to the target input will be strongest if applied at the decaying phase of the priming input, when the network is most receptive to new inputs. A speedup is achieved in comparison to a control experiment, where the first input is not applied at all; in the latter case, at the moment the target pattern is applied, the network is typically in a concept belonging to a different class. When the target input is applied only after the decaying phase of the priming input, the network's state has already undergone a transition, and the 'priming' experiment turns into a control experiment. Hence, although monotonically decreasing initially, the RT will eventually increase with delays larger than the length of the decaying phase.

5 Transitions in a damaged network

We now study the behavior of the network encompassing both semantic and episodic associations. Episodic associations are generated by an additional increment in the values of the original synaptic connections J_{ij}

(Eq. 9), according to:

$$\hat{J}_{ij} = J_{ij} + \kappa\left(\sum_{\mu=1}^{L_1}\sum_{\nu=1}^{L_2}(\xi_i^{\mu+1,\nu}-p)(\xi_j^{\mu\nu}-p) + (\xi_i^{\mu\nu}-p)(\xi_j^{\mu+1,\nu}-p)\right) \quad (9)$$

where $\mu = L_1 + 1$ is to be identified with $\mu = 1$. The two terms in the above equation represent symmetric pointers (Horn and Usher 1989) across classes, namely relating class μ with class $\mu + 1$ and vice versa, for the specific patterns ν inside classes μ and $\mu + 1$. According to this choice, the total number of semantic associations equals the number of episodic associations. The pointer strength κ is determined so that in an undamaged network 'semantic' and 'episodic' transitions occur with the same probability. As an approximation we choose κ equal to the incoming field at a neuron belonging to a nonactivated pattern of the same class as the currently activated pattern, i.e. $\kappa = (p_2 - p)/(1 - p)$.

The 'normal', undamaged network, has an initial level of semantic and episodic transitions that is higher than the level of the random transitions. Taking this as a starting configuration, we have examined the performance of the network when 'pathological' changes are inflicted. A random deletion of some fraction of the neurons leads to a concomitant decrease in both the semantic and episodic transitions to noise level, without any advantage to any specific type of transitions. A similar pattern was obtained when synaptic connections were randomly deleted. However, when deleting only neurons with a relatively large connectivity tree, i.e., whose sum of excitatory connections (which are both input and output connections) is large, an interesting pattern, shown in Fig. 5, is revealed: As neuronal deletion proceeds, the fraction of semantic transitions out of the total number of transitions occurring in the network actually rises.

The design of the latter simulation has been motivated by some recent neuroanantomical reports considering neuronal degeneration in Alzheimer's disease (AD). Although it is conventionally claimed that AD is accompanied by a considerable reduction of the neuronal mass, it has been recently reported that neuronal loss in AD is primarily limited to a specific subpopulation of large neurons (Terry et al. 1981; Hyman et al. 1984). We have interpreted this data in accordance with the work of Bok (1959) and Swindale et al. (1981), which have shown that the volume of the cell body is positively correlated with size of its connectivity tree, where the latter term includes both the neuron's axonal and dendritic trees. Neuropsychological tests, on the other hand, support the notion that in AD patients the performance of the semantic memory may severely decrease while some of the episodic memory capacities are still maintained (Granholm and Butters 1988; Salmon et al. 1988). Hence, the results of the second experiment show an interesting qualitative resemblance to the latter reports on AD. Obviously, as the disease advances, and neuronal degeneration continues, all memory capacities are severely damaged, which is known to be the clinical hallmark of AD (Adams and Victor 1989).

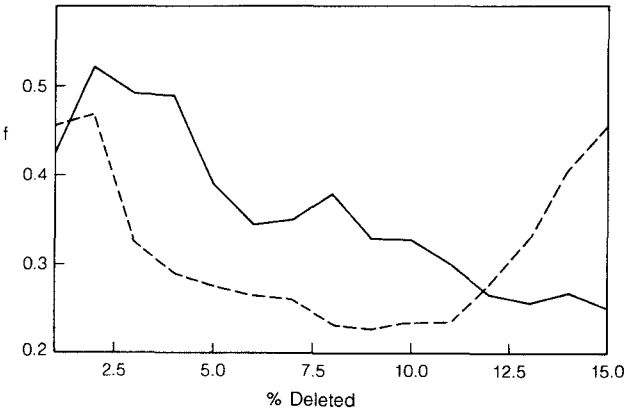

Fig. 5. Semantic (*full curve*) and episodic (*dashed curve*) transitions when large neurons are deleted. The fraction of semantic and episodic transitions is shown as function of the fraction of deleted neurons

It should be noted that the rise in the fraction of episodic transitions is maintained only for a short period (assuming that neuronal deletion continues at the same rate). The absolute number of episodic transitions actually constantly decreases. The most evident result obtained in our model (observed with all kinds of damage experimented) is the absolute reduction of the number of transitions occurring as the level of pathological damage is increased. This reduction may account for the paucity of speech and thought production observed in advanced stages of AD (Adams and Victor 1989).

6 Discussion

As concluded by Gröbler et al. (1991), it is of prime interest to investigate how a model for dynamical activation of memory may be supplemented by a neural level realization. At the conceptual level, the model presented here is defined by a network of memorized items segregated into hierarchical semantic classes, that are formed by proximity relations in a metric space defined by their attributes. The memorized concepts are linked by both semantic and episodic associations. At the dynamical level the model is defined as a TANN. On a short time scale (due to nonlinear competition leading to seriality), the model dynamics are charcterized by convergence to the memory patterns (on the basis of attribute similarity). On a longer time scale, transitions among the various memory items take place. The transition processes itself is stochastic, but guided by the semantic and episodic associations. As shown, when the network is in an attractor state, it is in high overlap with only one memory pattern. *Spreading of activation* hence takes place in our model during the network's transitions, when the network's state has considerable overlap with several memory patterns.

Since the TANN realization assigns a specific significance to attributes, not only in defining the semantic space, but also in determining the dynamics of the network, it seems quite difficult (if not impossible) to

present other realizations (e.g., a graph-like symbolic architecture) of our model that will implement all its conceptual properties. The two kinds of associations are realized in an inherently distinct manner, which cannot be simply mimicked by a graph-like symbolic realization. Thus, as has already been advocated (Smolensky 1986), the neural realization is not just a trivial mapping of the conceptual model.

We should view the model as simplified metaphor of the workings of the brain, in semantic processing. As shown in the Appendix the number of concepts in a class is essentially limited by the encoding probability ($L_2 < 1/p_2$). A more elaborate model for semantic processing should also accommodate for concepts having properties of different levels of significance, since out of all the attributes composing a concept some properties are more essential to the concept than others (Smith et al. 1974). Such a model should also account for the empirical distribution of properties shared by several concepts found by Rosh and Mervis (1975), and for higher order correlations between properties and concepts.

The model presented obviously addresses only a few of the mechanisms involved in thought processes, especially regarding the conscious-unconscious dichotomy. For example, a more elaborate model of these issues should account for phenomena such as subliminal priming (Marcel 1983), in which the RT to a target is facilitated by a weak stimulus that does not reach the subject's awareness. It seems that in this respect our network reaches a limitation inherent to ANNs; either the competition between memories is strong enough and then a subliminal stimuli will be quickly suppressed and thus ineffective, or the competition is so weak that the network's state will be most of the time spread in a mixture of memory states. Thus it may be necesary to supplement any ANN (or TANN) cognitive model by some external mechanisms modulating the degree of competition in the network. The existence of such external 'attentional' mechanisms has been advanced by Posner et al. (1988) on the basis of PET recording during attentional and nonattentional tasks.

The constraints enforced by the neural realization make any neural network modeling of cognitive phenomena a challenge. In this work, we have attempted to show that the analogy drawn between the neural-like architecture of the model's realization and biological memory systems, may have its own potential: After demonstrating that the main charcteristics of the dynamics of the conceptual model are preserved in its undamaged state, the neural model may be used to study the effects of distinct patterns of damage on its behavior. In general, such patterns of damage may be divided into structural versus dynamical malformations. We have examined only a subgroup of such structural patterns of damage, namely, neuronal deletion, synaptic deletion, and a more specified deletion of large neurons. However, the list of other structural changes that may be examined is fairly extensive; for example, one can make a distinction between the neuronal axonal and dendritic trees, and examine the behavior resulting from their damage separately. The neuronal population may be segregated to excitatory and inhibitory populations, and spatial organization of the neurons' connectivity may enable the investigation of the effects of spatially distinct patterns of damage. The investigation of dynamical pathogenetic changes depends too on the richness of the model. In principle, as the dynamical description of the neuron's input/output function gets more detailed, the effects of other dynamical parameters (such as membrane conductance changes, in addition to the noise and neural fatigue variations that were investigated here) can be examined.

From a biological point of view, it should be noted that recent advances in morphometric techniques have yielded new data on neuroanatomical changes that take place on the neuronal and synaptic level, in both normal aging and Alzheimer patients (e.g., Bertoni-Feddari et al. 1990; DeKosky and Scheff 1990). In addition to such structural findings, newly acquired data on possible pathological changes of the neuron's dynamics have also motivated recent neural models. In addition to the investigations of neurmodulation previously mentioned (Mamelak and Hobson 1989; Servan-Schreiber et al. 1990; Sutton et al. 1992), others have studied other effects of variations in neurotransmitter levels in biologically oriented neural models (King et al. 1984; Carpenter and Grossberg 1990; Hasselmo and Bower 1992). In spite of the inherernt difficulties involved in examining the microstructure of memory with existing experimental techniques, some important steps have already been done in recordings from the inferotemporal cortex of the monkey (Fuster 1990; Miller et al. 1991; Sakay and Miyashita 1991; Tanaka et al. 1991), and from the hippocampus and amygala of humans under clinical surgery (Heit et al. 1988). Surely this is only a start, but in the future neural models may indeed have an important role in studying the relation between microscopic neuropathological changes and macroscopic clinical phenomenology.

Acknowledgements. We are grateful to David Horn and Michail Tsodyks for many helpful discussions. M.H. was supported by a grant of the German Federal Department of Research and Technology. M.U. was a recipient of a Bantrell postdoctoral fellowship.

Appendix: Determination of the parameters

a) Mean post-synaptic potentials (fields)

The mean value of the neural field $\langle h_i \rangle$ is calculated using (4) and (5), averaging over the distribution of $\xi_j^{\mu\nu}$ (Eqs. 1–3).

Consider for example a the post-synaptic field obtained when the network's state is identical to the first memory pattern $S_i = \xi_i^{11}$. Since $\langle S_i \rangle = p$, the last term in (5) cancels and the post-synpatic field of neuron i is

$$\langle h_i \rangle = \left\langle \frac{1}{p(1-p)N} \sum_{j=1}^{N} \sum_{\mu\nu} (\xi_i^{\mu\nu} - p)(\xi_j^{\mu\nu} - p) S_j \right\rangle$$

$$= \left\langle \frac{1}{p(1-p)N} \left(\sum_{j=1}^{N} (\xi_i^{11} - p)(\xi_j^{11} - p) S_j + \sum_{j=1}^{N} \sum_{\nu > 1} (\xi_i^{1\nu} - p)(\xi_j^{1\nu} - p) S_j + N \right) \right\rangle$$

where N is a noise term having a zero mean originating from the sum over the patterns in other classes than ξ^{11}, $N = \Sigma_{j=1}^{N} \Sigma_{\mu>1,\nu} (\xi_i^{\mu\nu} - p)(\xi_j^{\mu\nu} - p) S_j$. The other two terms are the signal, and a noise term of non-zero mean, generated by the contributions from the other patterns in the same class. Consequently:

$$\langle h_i \rangle = (\xi^{11} - p) + \frac{1}{p(1-p)} \sum_{\nu>1} (\xi_i^{1\nu} - p)(p_1 p_2^2 - p^2)$$

Depending on the neuron, i, the following fields are obtained:

1. For neurons belonging to the pattern $\xi_i^{11} = 1$ (i.e., neurons that are 'ones' in that pattern), the mean synaptic field is bounded by a minimal value of

(a) $h_{1a} = (1-p) + (L_2 - 1)(p_1 p_2^2 - p^2)/p(1-p)(-p)$, for neurons that do not belong to any other pattern in the class ($\xi_i^{1\nu} = 0$ for all ν), and by a maximal value of

(b) $h_{1b} = (1-p) + (L_2 - 1)(p_1 p_2^2 - p^2)/p(1-p)(1-p)$, for neurons that belong to all other patterns in the class ($\xi_i^{1\nu} = 1$ for all ν).

2. For neurons that do not belong to the pattern ($\xi_i^{11} = 0$), the field is bounded by a minimal value of

(a) $h_{2a} = (-p) + (L_2 - 1)(p_2 p_2^2 - p^2)/p(1-p)(-p)$, for neurons that do not belong to any pattern in the class ($\xi_i^{1\nu} = 0$), for all ν), and by a maximal value of

(b) $h_{2b} = (-p) + (L_2 - 1)(p_1 p_2^2 - p^2)/p(1-p)(-p)$, for neurons that belong to all the patterns in the class except, ξ^{11}, ($\xi_i^{1\nu} = 1$, for all $\nu \neq 1$).

In order to obtain a stability of the patterns, the parameter should be chosen so that the values of the field h_i obtained in case 1*a* and 2*b*, above, are separable, i.e.,

$$(1-p) + (L_2 - 1)\frac{p_1 p_2^2 - p^2}{p(1-p)}(-p) \gg (-p) + (L_2 - 1)\frac{p_1 p_2^2 - p^2}{p(1-p)}(1-p)$$

If follows that $L_2 - 1 \ll (1-p)/p_2 - p$. Using the sparseness of the encoding ($p_1, p_2 \ll 1$), a constraint for the number of concepts in a class L_2 is obtained; $L_2 < p_2^{-1}$ (Tsodyks 1990). However, no restriction is imposed on p_1 and L_1. In the numerical simulations a lower bound for the activities p is imposed by the requirement that pN should not be too small. For $N = 500$, we have thus chosen $p_1 = 1/3$, $p_2 = 1/4$ and $1/6$, which obey these constraints.

b) Thresholds

The threshold's value should lie between the values of the neural field h_{1a} and h_{2b}, in order to guarantee firing for neurons encoding properties that belong to the pattern but not to the class, and inactivity for neurons that encode properties that belong to the class but not to the specific pattern. The 'optimal' total threshold $\theta = \theta^0 + \theta_i(t)$, that is the medium value of the two fields is:

$$\theta = (\tfrac{1}{2} - p)\left(1 + \frac{p_2 - p}{1-p}(L_2 - 1)\right).$$

In order to obtain semantic transitions, the threshold θ is considered a dynamic variable, whose variation range lies between the value of the fields in case 1*a*, and 2*b*. In our simulations, we have chosen $\theta_0 = 0.35$, the value of which is higher than the value of the field $h_{2b} = 0.25$(for $p_1 = 1/3$ and $p_2 = 1/4$). The variable threshold $\theta_i(t) > 0$ has a maximum of $\theta_{\max} = 0.4$. Thus the maximal total threshold $\theta = 0.75$ is close to the field value $h_{1a} = 0.9$ (for $p_1 = 1/3$ and $p_2 = 1/4$). Thus, when the threshold gets close to its maximal value, some of the neurons that encode properties specific to the pattern but not to its class (case 1*a*) will be deactivated by the noise. Neurons that encode properties belonging both to a pattern and to its class have higher fields (case 1*b*) and are therefore less affected by the fatigue. Therefore the resulting transitions are biased towards the semantic class. Increasing the noise T or the maximal threshold value, has hence the effect of diminishing the semantic bias, as shown in text.

c) Strengths of the pointers

Incorporating episodic transitions into our model, the mean value $\langle h_i \rangle$ for $S_i = \xi_i^{11}$ (using Eqs. 4, 5, 9) is

$$\langle h_i^{(\kappa)} \rangle = \xi_i^{11} - p + \frac{p_2 - p}{1-p} \sum_{\nu>1} (\xi_i^{1\nu} - p) + \kappa \sum_{\mu>1} (\xi_i^{\mu 1} - p)$$

In order to achieve similar transition rates for both semantic and episodic transitions, the parameter κ has to be chosen so that the mean values of the second and the third term are equal, which gives $\kappa = p_2 - p/1 - p$. For $p_2 = 1/4$ and $p_2 = 1/6$ we have $\kappa = 2/11$ and $\kappa = 2/17$, respectively.

References

Abbott LF (1990) Modulation of function and gated learning in a network memory. Proc Natl Acad Sci USA 87:9241–9245

Adams RD, Victor M (1989) Principles of neurology. McGraw-Hill, New York

Amit DJ (1989) Modeling brain function: the world of attractor neural networks. Cambridge University Press, Cambridge

Anderson JR (1976) Language, memory, and thought, Erlbaum, Hillsdale

Andersson JR (1985) Cognitive psychology and its implications. W. H. Freeman and Company, New York

Bertoni-Freddari C, Fattoretti P, Casoli T, Meier-Ruge W, Ulrich J (1990) Morphological adaptive response of the synaptic junctional zones in the human dentate gyrus during aging and Alzheimer's disease. Brain Res 517:69–75

Bok ST (1959) Histonomy of the cerebral cortex. Elsevier, Amsterdam.

Carpenter GA, Grossberg S (1990) Art 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. Nerual Networks 3:129–152

Collins A, Loftus E (1975) A spreading activation theory of semantic processing. Psychol Rev 82:407–428

Collins AM, Quillian MR (1969) Retrieval from semantic memory. J Verb Learn Behav 8:240–247

DeKosky ST, Scheff SW (1990) Synapse loss in frontal loss in frontal cortex biopsies in Alzheimer's diease: Correlation with cognitive severity. Ann Neurol 27(5):457–464

Fahlman S (1981) Representing implicit knowledge. In: Hinton GE, Anderson JA (eds) Parallel models of associative memory. Erlbaum, Hillsdale

Feigelman MW, Ioffe LB (1987) The augmented models of associative memory asymmetric interaction and hierarchy of patterns. Int J Mod Phys B 1:51–68

Fuster J (1990) Inferotemporal units in selective attention and short term memory. J Neurophysiol 64:681–697

Georgopolous AP, Kettner RE, Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three dimensional space. II. Coding of the direction of movement by a neural population. J Neurosci 8: 2928–2937

Granholm E, Butters N (1988) Associative encoding and retrieval in Alzheimer's and Huntington's disease. Brain-Cogn 7(3):335–347.

Gröbler T, Marton P, Erdi P (1991) On the dynamic organization of memory. Biol Cybern 65:73–79

Gutfreund H (1988) Neural networks with hierarchically correlated patterns. Phys Rev A 37:570–577

Hasselmo ME, Bower JM (1992) Chlorinergic modulation of cortical memory function. J Neurophysiol 67:1230–1246

Heit G, Smith ME, Halgren E (1988) Neural encoding of individual words and faces by the human hippocampus and amygdala. Nature 333:773–775

Herrmann M, Tsodyks M. V. (1991) Pattern hierarchy destruction in nonlinear neural networks. Workshop on Neural Networks in Biology and High Energy Physics, Isola d'Elba, 1991

Hinton GE (1981) Implementing semantic networks in parallel hardware. In: Hinton GE, Anderson JA (eds) Parallel models of associative memory. Erlbaum, Hillsdale

Hoffman RE (1987) Schizophrenia-mania dichotomy. Arch Gen Psychiatry 44:178–191

Hoffman RE, Dobscha S (1989) Schizophrenia Bull 15(3):477–489

Horn D, Usher M (1989) Neural networks with dynamic thresholds. Phys Rev A 40:1036–1044

Horn D, Usher M (1990) Excitatory-inhibitory networks with dynamical thresholds. Int J Neural Syst 1:249–257

Hyman BT, Van Hoesen GW, Damasio AR, Barnes CL (1984) Science 225:1168–1170
Kihlstrom JF (1987) The cognitive unconscious. Science 237:1145–1152
King R, Barchas JD, Huberman BA (1984) Chaotic behavior in dopamine neurodynamics. Proc Natl Acad Sci USA 81:1244–1247
Kleinfeld D (1986) Sequential state generation by models of neural networks. Proc Natl Acad Sci USA 83:9469–9473
Mamelak AD, Hobson JA (1989) Dream bizarreness as the cognitive correlate of altered neuronal behavior in REM sleep. J Cogn Neurosci 1(3):201–221
Marcel AJ (1983) Conscious and unconscious perception: an approach to the relations between phenomenal and perceptual processes. Cognit Psych 15:238–300
Miller EK, Li L, Desdimone D (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254:1377–1379
Posner MI, Petersen SE, Fox PT, Raichle ME (1988) Localization of cognitive operations in the human brain. Science 240:1627–1631
Ratcliff R, McKoon G (1981) Does activation really spread? Psychol Rev 88:454–462
Ritter H, Kohonen T (1989) Self-organizing semantic maps. Biol Cybern 61:241–254
Rosh E, Mervis C (1975) Family resemblances: studies in the internal structure of categories. Cogn Psych 7: 573–605
Rumelhart DE, McClelland J (1986) Parallel distributed processing. MIT Press, Cambridge
Sakay K, Miyashita Y (1990) Neural organization of the long term memory of pair associates. Nature 354:152–159
Salmon DP, Shimamura AP, Butters N, Smith S (1988) Lexical and semantic priming deficits in patients with Alzheimer's disease. J Clin Exp Neuropsychol 10(4):477–94
Schacter DL (1989) Memory. In: Posner MI (eds) Foundations of cognitive science. MIT Press, Cambridge, pp 683–708
Servan-Schrieber D, Printz H, Cohen JD (1990) A network model of cathecholamine effects: gain, signal-to-noise ratio, and behavior. Science 249:892–895
Skarda CA, Freeman WJ (1987), Does the brain make chaos in order to make sense of the world. Behav Brain Sci 10:161–165
Smith E, Shoeben E, Rips LJ (1974) Structure and process in semantic memory: a featural model for semantic decisions. Psych Rev 81:214–232
Smolensky P (1986) Neural and conceptual interpretation of PDP models. In: Rumelhard DE, McClelland J (eds) Parallel distributed processing. MIT Press, Cambridge, pp 390–431
Sompolinsky H, Kanter I (1986) Temporal associations in assymetric networks. Phys Rev Lett 57:2861–2864
Squire LR (1982) The neuropsychology of human memory. Ann Rev Neurosci 5:241–273
Sutton JP, Mamelak AN, Hobson JA (1992) Modeling states of waking and sleeping. Psychiatr Ann 22(2):1–7
Swindale NV, Vital-Durand F, Blakemore C (1981) Recovery from monocular deprivation in the monkey. Reversal of anatomical effects in the visual cortex. Proc R Soc Lond Sec B. 213:435–450
Tanaka K, Saito Y (1991) Coding visual images of objects in the inferotemproal cortex of macaque monkey. J Neurophysiol 66:170–189
Terry RD, Peck A, Teresa R, Schechter R, Hyoroupian DS (1981) Ann Neurol 10:184–192
Treisman A, Gelade G (1980). A feature-integration theory of attention. Cogn Psychol 12:97–136
Tsodyks MV (1990) Hierarchical associative memory in neural networks with low activity level. Mod Phys Lett B 4:259–265
Tulving E (1985) How many memory systems are there? Am Psychol 40:385–398