

A Connectionist Model of False Memories

Saskia van Dantzig (vandantzig@fsw.eur.nl)

Psychology Department, Erasmus University Rotterdam
Postbus 1738, 3000 DR Rotterdam, The Netherlands

Eric O. Postma (postma@cs.unimaas.nl)

Department of Computer Science, University of Maastricht
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract

We present a connectionist model of false memories called the Associative Self-Organizing Network (ASON) model. Four mechanisms underlying the Constructive Memory Framework (CMF) guide the design of the ASON model, a connectionist operationalisation of the CMF. Simulation studies of experiments in the DRM paradigm reveal the ASON model to exhibit false memories. In addition, the effects of Mean Backward Associative Strength and output order on the probability of false recall are simulated. We conclude that the ASON model is capable of simulating and explaining the main findings on false memories.

Introduction

Memory is fallible. Every day people are confronted with the shortcomings of their memory, when forgetting things such as -for example- the phone number of a good friend, the title of a book, or the location of their car keys. Memory can also fail in another way; instead of forgetting things that did happen, people may remember events that never took place. These memories can be just as realistic as memories of real events. Such memories of never-happened episodes are called *commission errors* or *false memories*. False memories may occur in different situations and their severity can range from attributing a memory to the wrong source to confabulating a complete event (Parkin, 1997).

Various studies suggest that false memories are not simply random errors (Gallo & Roediger, 2002; Schwartz et al., 1998). Instead, they appear to be an inevitable consequence of the dynamics of human memory (Schacter et al., 1998). False memories are considered to arise from the very same mechanisms that underlie veridical recall and recognition of true memories. More specifically, we hypothesize that false memories result from the way in which memory representations are stored, processed, and retrieved.

Our approach is to investigate the occurrence of false memories in a connectionist model called the Associative Self-Organizing Network (ASON) model. The ASON model is made up of two associatively connected self-organizing maps, for storing and representing stimuli and the contexts in which they occur. Although the scientific literature on false memories is abundant (e.g. Gallo & Roediger, 2002; Johnson et al., 1993; Schacter et al., 1998), to our knowledge, no connectionist model of false memories has yet been proposed.

The outline of the remainder of this paper is as follows. In the next section we discuss the theoretical background. Then, we present the Associative Self-Organizing Network as a model of false memories. In addition, three simulations are described. Finally, we discuss the results and conclude upon the approach.

Theoretical Background

The common view of memory is that of a (re)constructive process (Roediger & McDermott, 1995; Schacter et al., 1998). This means that memories, rather than being literal reproductions of past events, are considered to be reconstructions that are susceptible to a variety of distorting factors. In this view, memories are distorted by schemes, attribution processes, prior knowledge, assumptions, and so forth. This makes it almost impossible to draw a clear boundary between true and false memories in real life situations. For this reason, our study focuses solely on false memories occurring in the experimental setting of the Deese-Roediger-McDermott (DRM) paradigm. A false memory is formalized as a recollection of a stimulus that is ascribed to the experimental context, whereas it was *not* presented during the experiment. Below, we describe the DRM paradigm in more detail.

The DRM Paradigm

In order to investigate false memories experimentally, Roediger and McDermott (1995) developed the DRM paradigm, which was a variation of a design originally used by Deese (1959). The experimental set up is as follows. Subjects are presented with lists of twelve or fifteen words that are the strongest associates of a “critical lure”; a target word which is not presented. Immediately following the presentation of a list, subjects are instructed to recall as many of the list items as possible and to mention only those words of which they are certain that they appeared on the list. Despite this instruction, subjects are about equally likely to recall the critical lure as the other items on the list (Roediger & McDermott, 1995). After completion of the experiment, which usually involves the presentation of multiple lists, recognition performance of items on all the lists is tested. It is found that subjects identify the critical lure as being a list item as often as or more often than words that were actually presented (Roediger & McDermott,

1995). These results have been widely replicated, using various lists and different variations on the basic paradigm.

The propensity to elicit false recall and false recognition of the critical item varies widely with the type of list used. Roediger et al. (2001) investigated the causes of this variability and found that the strongest predictor of false recall of the critical lure was a variable called Mean Backward Associative Strength (MBAS). MBAS is defined as the average probability that a list item elicits the critical item as its associate. Roediger et al. found that MBAS correlates positively with both false recall ($r = +.70$) and false recognition ($r = +.43$) of the critical lure.

The ASON model is inspired by the Constructive Memory Framework of Schacter et al. (1998). In the following section this framework is discussed in detail.

The Constructive Memory Framework

Many different theories exist that address the topics of memory formation, source monitoring or reality monitoring and false memories (e.g. Gallo & Roediger, 2002; Johnson et al., 1993; Reyna & Brainerd, 1995). The general assumption underlying these different theories is that memory is constructive. This is also the central assumption of the Constructive Memory Framework (CMF) (Schacter et al., 1998). CMF proposes four mechanisms that are involved in a constructive memory system.

First, according to CMF, episodic memories can be viewed as patterns of features, with different features representing different aspects of the episode. The constituent features of a memory representation are distributed widely across different parts of the brain. Forming an episodic memory involves binding together an arbitrary configuration of information from different sources (visual, auditory, affective, semantic etcetera) about a specific episode into a unitary whole (O'Reilly & Rudy, 2001; Rolls & Treves, 1998; Schacter et al., 1998). This process is called *feature binding*.

Second, each episode activates a unique representation that can easily be discriminated from memories of similar events. Even if different memories overlap extensively, the memory system is able to retrieve the unique characteristics of each particular episode, rather than retaining only the general similarities or gist (Reyna & Brainerd, 1995). This requires a process called *pattern separation* (Schacter et al., 1998).

Third, retrieval of memories involves a process of *pattern completion*. At retrieval, a small part of the original memory is used as a retrieval cue. The subset of features representing this part of the memory is activated. Activation spreads from the activated features to the rest of the constituent features that represent that experience, and the complete memory is reconstructed.

Fourth, once a memory is reconstructed, it must be decided whether the retrieved information constitutes a real memory or is derived from internally generated information, such as thoughts or fantasies. This process is called reality monitoring. Source monitoring is a broader concept and

refers to determining the source of a retrieved memory. According to Source Monitoring Theory (Johnson et al., 1993), memories from different sources have different qualitative characteristics. Source monitoring decisions capitalize on these differences. When the source monitoring mechanism fails, source amnesia occurs. One is then able to remember specific information, but unable to recall the source of this memory.

The four mechanisms of the Constructive Memory Framework lead to the notion that false memories result from a combination of two factors: (1) memories from different sources (e.g. internal and external) may form overlapping representations, and (2) the source monitoring mechanism fails to distinguish between those representations.

The activation/monitoring framework, (Gallo & Roediger, 2002; Roediger et al., 2001) explains variations in the probability of false remembering in the DRM paradigm in terms of the two factors. According to this framework, two processes, activation and monitoring, take place during the encoding and retrieval of memories. Although activation occurs mostly at the encoding stage and monitoring mostly at the retrieval stage, both processes are at work during both encoding and retrieval. The activation/monitoring framework assumes that the presentation of some items can activate entire knowledge structures or schemata. As a consequence, non-presented items can be activated because they are strongly associated with the presented items (i.e., they are part of the same knowledge structure). The activation may be the result of conscious, deliberate association, or of automatic and unconscious spreading activation. In the case of the DRM paradigm, activation spreads from the list items to related or associated concepts. The critical lure receives much activation because this item is strongly associated to each of the presented list items. This assumption is supported by the high correlation between MBAS and false remembering of the critical lure. The stronger the association between the list items and the critical word, the stronger the activation of this critical word due to automatic or deliberate spreading of activation.

Summarizing, false remembering of the critical lure occurs when the monitoring process fails to correctly attribute its activation to an internal source and the critical lure is falsely ascribed to the learning context. This monitoring process is analogous to the source monitoring or reality monitoring mechanism proposed by Johnson et al. (1993).

Implementing CMF in a Connectionist Network

The CMF acted as a guideline for the design of the ASON model. The four mechanisms of the CMF translate into the following four desired abilities of the ASON model.

- (1) Ability to form episodic memories, whereby each episode leaves a unique, distinctive trace that is easily distinguishable from memories of similar episodes (i.e., demonstrate feature binding and pattern separation).

- (2) Ability to retrieve or reconstruct a complete representation when cued with only a small part of the original memory (i.e., exhibit pattern completion).
- (3) Ability to spread activation among related or associated concepts.
- (4) Ability to monitor memory using a mechanism that decides upon the trueness of each retrieved representation.

We incorporate the four abilities in the ASON model as follows.

(1/2) Feature binding and pattern completion. Feature binding is accomplished by using an associative network, or more specifically, an auto-associator. An auto-associator typically consists of one fully-connected layer. The network's task usually is to produce an output that is similar to its input. When an input pattern is presented, the network's connection weights are changed according to a Hebbian learning algorithm. Connections between simultaneously active neurons are strengthened, whereas connections between non co-active neurons are weakened. In this way, the network is able to associate co-occurring input elements. In addition, the auto-associator is able to completely reconstruct a stored pattern, when provided with only a small part of that pattern. In other words, it can also perform pattern completion (McLeod et al., 1998).

(1/3) Pattern separation and spreading activation. In an auto-associative network, pattern separation is obtained by using sparsely distributed representations. A competitive network can be used to transform densely distributed input patterns into more sparse, separated patterns which can be processed by an auto-associator without suffering from interference. A specific kind of competitive network is the Kohonen network or self-organizing network (Haykin, 1999). For our purposes, the self-organizing network has two important advantages over a standard competitive network. First, the self-organizing network creates a topological map of the input space (Haykin, 1999). A distributed, multidimensional input is transformed into a localist representation. The self-organizing principle ensures that the information regarding relations or similarities among input patterns is not lost in this transformation. By creating a topological map of the input space, the similarity between two input patterns is reflected in the lateral distance between the two neurons representing them. This is a biologically plausible way of representing information. There is evidence that at least lower level sensory representations are organized topologically (Haykin, 1999). However, it is still uncertain whether semantic information in higher association areas is represented in a topological way as well. A second important characteristic of a self-organizing network is that there is spreading of activation among neighboring neurons. When a specific neuron in the network is excited, activation spreads to its neighbors. The degree of spreading activation is a function of the distance between the excited neuron and its neighbor. The nearest neighbors receive the most activation, and activation

decreases with increasing distance. Since the neighbors of the winning neuron represent concepts resembling the input pattern, there is spreading activation between related concepts. In this way the network resembles a semantic, or conceptual map. It is generally assumed that much of our knowledge is indeed stored in the form of semantic maps or knowledge structures.

(4) Memory monitoring. A memory monitor mechanism may be implemented in the form of a module that modulates the response thresholds or connection weights of neurons in the associative layer.

In the next section the incorporation of the ASON model is described in detail.

The Associative Self-Organizing Network

The ASON model, shown in figure 1, receives two different types of input; *context* input and *stimulus* input. Input is first processed by the input/output layer of the model. This layer is made up of two unconnected parts. One part processes contextual information, the other part deals with stimulus information. Both parts of the input/output layer have I neurons. The input of the model is formed by multidimensional binary patterns. In those patterns, each bit represents the presence or absence of a specific feature by which the stimulus (item) or context is characterized. The input patterns therefore reflect conceptual representations of different stimuli and contexts.

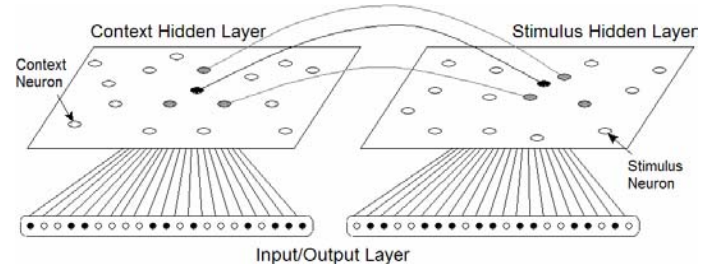


Figure 1: Schematic drawing of the Associative Self-Organizing Network. Each input pattern corresponds to one winning neuron in the hidden layer. Simultaneous presentation of a stimulus input and a context input causes an increase in the connection strength between the hidden neurons representing that stimulus and context, respectively.

Information is propagated from the input/output layer to the hidden layer. The hidden layer consists of two self-organizing maps. These maps are organized as two-dimensional lattices, each having $N \times N$ neurons. The part of the hidden layer that represents stimuli is henceforth called the *stimulus hidden layer*. The neurons making up this layer are called *stimulus neurons*. The other part of the hidden layer is called the *context hidden layer* and the constituent neurons are called *context neurons*. Both hidden layers are fully connected to each other via two-directional modifiable associative connections.

The Four Processing Stages

The processing of information in the ASON model proceeds in four stages: (i) the initialization stage, (ii) the topological-mapping stage, (iii) the learning stage, and (iv) the performance stage. Below, we discuss each of these stages in detail.

In the initialization stage, the connection weights between the input/output layer and the hidden layers, and those between both hidden layers are set to small random values.

In the topological-mapping stage, contexts and stimuli are presented to the input layer of the network and, using the Kohonen learning algorithm or SOM algorithm (Haykin, 1999), a topological organization in the hidden layers is created: semantically related concepts (overlapping input patterns) are represented by neurons that lie close to one another in the two-dimensional grid that makes up the hidden layer. In addition, associations are formed between stimuli and contexts. Whenever a particular stimulus co-occurs with a particular context, there is simultaneous activation of the winning context neuron and the winning stimulus neuron. Following an associative learning algorithm, the (associative) connection between these two hidden neurons is strengthened. Simply said, the stimulus is coupled to the context.

The learning stage simulates the learning phase of the DRM task. It refers to the presentation of the list items. During this stage, a number of stimuli are presented in one specific context -the learning context- and associations between the presented stimuli (the list items) and this context are formed. Due to spreading of activation, not only the connections between the winning context neuron and the winning stimulus neurons are strengthened, but also those between the context neuron and the neighbors of the winning stimulus neuron. The connections between the context neuron and even further neighbors of the winning stimulus neuron are actually decreased.

During the performance stage, the network can either perform a recall task or a recognition task. When performing a recall task, a context input is presented to the network as a recall cue. The winning context neuron in the hidden layer is determined and activation is propagated forwards through the associative connections towards the stimulus hidden layer. The stimulus neuron that is most strongly associated to the winning context neuron is activated and propagates its activation to the stimulus input/output layer. The weights of the connections between the winning stimulus neuron and the input/output layer have changed during the topological-mapping stage so that they have come to resemble the input pattern to which this neuron responds most strongly. Therefore, propagating activation through these connections will result in an output that resembles the original input pattern to a large degree. In other words, reconstruction of the stimulus that is most strongly associated with the presented context takes place. Subsequently, the connection between the winning context neuron and the activated stimulus neuron is 'blocked', the stimulus neuron with the second-strongest association to the

context is determined and the next stimulus is recalled.

Most of the time, the stimulus recalled is one that was actually presented during the learning stage (a list item). Occasionally, however, the network recalls a stimulus that has not been presented. In other words, it has false memories. Clearly, false memories occur whenever there exists a strong association between the non-presented stimulus and the context, caused by spreading activation.

When performing a recognition task, the network is presented with a number of stimuli, both list items and a number of non-presented distractors (including the critical item). Based on the stimulus input, the winning stimulus neuron in the hidden layer is determined. The strength of the association between this winning stimulus neuron and the learning context is determined. If the strength exceeds a certain threshold, the stimulus is marked as a target and as a distractor otherwise. The decision whether to accept or reject a retrieved item is based on the strength of its association to the learning context. Raising the threshold reduces the probability of falsely recognizing the critical item, but it also decreases the hit rate. On the other hand, lowering the threshold leads to more hits, but also to more false alarms. This process is a formalization of the memory monitoring or source monitoring mechanism in various theories of memory (Gallo & Roediger, 2002; Johnson et al., 1993; Schacter et al., 1998).

To evaluate the ability of the ASON model to exhibit the false-memory performance as observed in the DRM paradigm, we performed a number of simulations that are described in the following section.

Simulations

Our simulations focus on three aspects of false memories in the DRM paradigm: the DRM effect, the role of association strength and the output order effect. All simulations were performed with the following parameter values: $I = 30$ and $N = 10$. The results reported do not depend critically on these choices.

The DRM effect The simulation of the DRM effect has two conditions; the DRM condition and a control condition. In the DRM condition, the network learns six list items that are semantically related to the critical lure. Their input patterns closely resemble the input pattern of the critical lure. In the control condition, the six list items are randomly chosen from the input set. After learning the six list items, the network performs a recall task. The results of the simulation are shown in figure 2. As is evident from the graph, the probability of recalling the critical lure is much higher in the DRM condition ($P = 0.65$) than in the control condition ($P = 0.05$), but it is lower than the average recall rate of the list items ($P = 0.78$). Hence, the ASON model simulates the DRM effect faithfully.

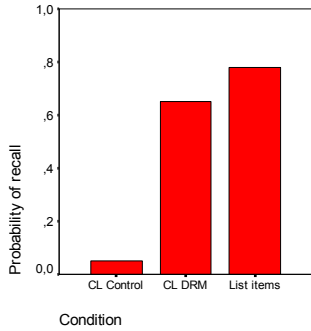


Figure 2: Probability of recalling the list items and the critical lure in the DRM (CL DRM) and control (CL Control) conditions.

The effect of association False recall and recognition of the critical lure is affected by a number of factors. As stated in the introduction, the most important factor is the Mean Backward Associative Strength (MBAS). A stronger association between the list items and the critical lure is correlated with stronger false recognition and false recall effects (Roediger et al., 2001). In the Associative Self-Organizing Network, spreading activation from the list items to the critical lure causes false recall and recognition of the latter. It is important to realize that in the ASON model concepts are related semantically instead of associatively. In contrast to what is proposed by the activation/monitoring framework, activation spreads along semantic relations rather than along associative connections. However, if we disregard this difference, we can define the Backward Associative Strength between two concepts as the lateral distance between the winning neurons that represent these concepts. The smaller the distance, the more related the two concepts are. In the network, the degree of spreading activation is a function of the distance between the excited neuron and its neighbor. Consequently, the smaller the average lateral distance between the list items and the critical lure, the stronger the activation of this critical item due to spreading activation from the list items will be. This stronger activation leads to a stronger association of the critical lure to the context, and therefore to an increasing likelihood of falsely recalling the critical lure. Figure 3a shows the results of a simulation in which the average lateral distance from the list items to the critical lure is varied. As can be seen, the probability of recalling the critical lure decreases sharply with increasing distance. The correlation between lateral distance and probability of recalling the critical lure is -0.76 . We compare our results with the results from a multiple regression analysis done by Roediger et al. (2001) where MBAS was found to be the strongest predictor of false recall of the critical lure (with the correlation between MBAS and probability of recalling the critical lure being $+0.73$). Figure 3b shows the probability of recalling the critical lure as a function of MBAS, as found in the study of Roediger et al. (2001). Clearly, the results of the ASON model agree very well with those of Roediger et al.

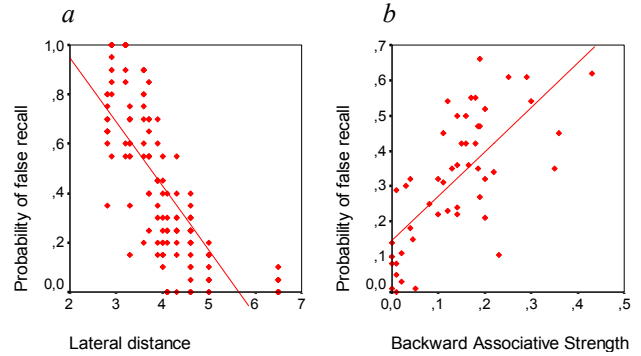


Figure 3: The probability of recalling the critical lure as a function of (a) average lateral distance between list items and the critical lure: $r = -0.76$, and (b) MBAS: $r = +0.73$.

The effect of output order In the third simulation we investigated the effect of output order on the probability of false recall. The *output order effect* (Schwartz et al., 1998) refers to the finding that the probability of a false memory increases with the position of items in the recall sequence.

The output order effect can be explained by the variation in association strength of presented and non-presented stimuli. False memories occur when non-presented stimuli, become strongly connected to the learning context through the processes of spreading activation and association. The association strength of those stimuli to the context is usually smaller than that of the most strongly associated targets, but larger than that of the most weakly associated targets. Since memories are generated in the order of their association strength, the probability that a false memory is generated increases with the position in the recall sequence. In our third simulation, the network performed a simple recall task, rather than a DRM task. The network learned twenty stimuli in a single context. Afterwards it performed a recall task. As can be seen in figure 4a, the probability of a false memory is largest in the last quartile of the output. Figure 4b shows the results of Schwartz et al. (1998), in which subjects performed a similar task. Evidently, our results have a striking similarity to the experimental results of Schwartz et al.

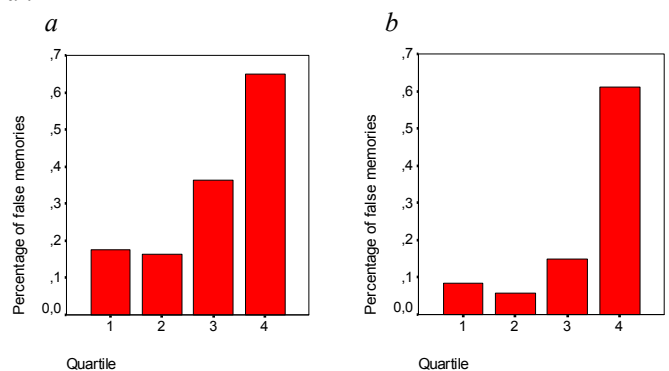


Figure 4: Number of false positives as a function of output order. Results of (a) our simulations, and (b) Schwartz et al. (1998).

Discussion and Conclusion

The ASON model demonstrates how the essential features of a constructive memory system, as put forward by CMF, can be translated into a connectionist model. Specifically, the ASON model incorporates the encoding processes of feature binding and pattern separation, as well as the retrieval processes of pattern completion and memory monitoring. In addition, it explains how spreading activation leads to high false memory scores for the critical lure in the DRM paradigm. The remaining question is to what degree the model's architecture resembles that of brain structures that are involved in the processes of storing, retrieving and monitoring of memory.

The brain structure that is considered to be responsible for the storage of episodic memories is the hippocampus (Rolls & Treves, 1998). The hippocampus is not thought to be the site of storage itself. Rather it is regarded as the mechanism that binds together the sensory features of a situation or episode to create a unitary representation of the experience. In other words, it is the structure that performs feature binding. The hippocampus receives, via the adjacent parahippocampal gyrus and entorhinal cortex, inputs from virtually all association areas in the neocortex. In addition, it gets input from the amygdala and from cholinergic and other regulatory systems (Rolls & Treves, 1998). It thus receives highly elaborated, multimodal information from various sensory pathways. Within the hippocampus, information is processed along a mainly unidirectional path, consisting of three major stages; the Dentate Gyrus (DG), the Cornu Ammonis 3 (CA3) and the Cornu Ammonis 1 (CA1). From CA1, backprojecting pathways lead via the subiculum and the entorhinal cortex back to the neocortex.

The hippocampus shares two essential characteristics with our model. First, there is a large degree of interconnectivity among neurons in the CA3 area of the hippocampus. This interconnectivity makes this area perfectly suited to perform auto-association. In fact, the idea that the CA3 area serves as an auto-associator that binds together the various elements of an episode is a core assumption in a number of computational models (O'Reilly & Rudy, 2001; Rolls & Treves, 1998). Second, the hippocampus receives a load of multimodal information from various cortical areas. The forward pathways to the hippocampus are thus characterized by strong convergence. It is hypothesized that these pathways, and the DG in particular, serve as a competitive network, transforming the widely distributed information in the cortex into more sparse, orthogonal and separated patterns that can be processed by the auto-associator without much interference (O'Reilly & Rudy, 2001).

Instead of a standard competitive network, the ASON model features a self-organizing map. The specific characteristics of this type of network, its ability to form a topological map of the input and spreading activation among neighboring neurons, can provide an explanation of false memories. Specifically, according to our model, false memories arise when activation spreads from the list items to the critical lure, causing a faulty association between this

non-presented item and the learning context. This explains how false memories occur in the DRM paradigm, and gives an account of the effect of MBAS on false recall of the critical lure.

By incorporating the four mechanisms of the CMF, the ASON model is able to simulate the occurrence of false memories in the DRM paradigm and the effects of MBAS and output order on the probability of false recall of the critical item. Furthermore, its architecture is compatible with that of the hippocampus, the brain area that is widely acknowledged as being involved in the storage and retrieval of episodic memories. We conclude that this connectionist operationalisation of the CMF is able to simulate and explain the main findings on false memories.

References

- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17-22.
- Gallo, D. A., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring. *Journal of Memory and Language*, *47*, 469-497.
- Haykin, S. (1999). *Neural networks, a comprehensive foundation* (2nd ed.). Upper Saddle River: Prentice-Hall.
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, *114*, 3-28.
- McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.
- O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311-345.
- Parkin, A. J. (1997). The neuropsychology of false memory. *Learning and Individual Differences*, *9*, 341-357.
- Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: some foundational issues. *Learning and Individual Differences*, *7*, 145-162.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 803-814.
- Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: a multiple regression analysis. *Psychonomic Bulletin and Review*, *8*, 385-407.
- Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.
- Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *48*, 289-318.
- Schwartz, B. L., Fisher, R. P., & Hebert, K. (1998). The relation of output order and commission errors in free recall and eyewitness accounts. *Memory*, *6*, 257-275.