NEUROIMAGING, BEHAVIORAL, AND COMPUTATIONAL INVESTIGATIONS OF MEMORY TARGETING

Sean Matthew Polyn

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF PSYCHOLOGY

MAY 2005

Abstract

The conceptual focus of this dissertation is the ability of humans to target episodic memories, i.e., to re-access past states of the world encoded by the memory system. I describe a framework for understanding this process that relies on the interaction of three cognitive systems in the brain: semantic memory, episodic memory, and a context maintenance system that acts to probe the episodic memory system. I build upon previous studies of these systems, in which each of these cognitive systems is mapped onto a particular anatomical region of the brain. Respectively, these areas are posterior cortex, medial temporal lobe, and prefrontal cortex. This framework is investigated in a series of studies of the free recall paradigm.

First, I describe a neuroimaging experiment in which I use pattern classification methods to track the second-by-second fluctuations of patterns of brain activity during free recall. The results of this experiment provide evidence of contextual reinstatement processes during the recall period. These results highlight the idea that context effects on memory organization can be carried out by any brain area that has both a relatively stable pattern of activity during the encoding process, and connections with medial temporal lobe brain regions.

Second, a behavioral study manipulates the accessibility of memories by changing encoding task midway through the study list, causing the sets of items encoded in the context of each task to become isolated from each other in memory. The results of this experiment suggest that patterns of task-related activity act as an effective context, uniting memory traces encoded within-task and isolating traces encoded under different tasks.

Finally, I describe a connectionist model of the interactions of these brain regions during free recall. I present a series of simulations that illustrate how the prefrontal system can manipulate the medial temporal memory system to flexibly extract previously encoded memories. The model also captures the "memory isolation" effects seen in the encoding task-shift experiment described above. A final set of simulations attempt to capture the pattern of deficits seen in elderly subjects in free recall (Kahana, Howard, Zaromb, & Wingfield, 2002).

Contents

Acknowledgements

I would like to thank my advisors, Kenneth Norman and Jonathan Cohen, for reasons too numerous to mention. I had much assistance in generating the results reported herein, and would like to thank Vaidehi Natu for tireless programming in the neuroimaging experiment, and Cara Buck for running the subjects in the behavioral experiment. Many thanks to Leigh Nystrom for sage advice in all domains, and to the members of the Cohen and Norman labs for valuable discussions and assistance. Many thanks, as well, to Marc Howard for providing figures and sharing ideas. Also, thanks to Sam Glucksberg and Robert Schapire for valuable comments and suggestions on prior versions of this document. I was supported by a National Research Service Award from the National Institute of Mental Health (MH070177-01).

Finally, I would like to dedicate this dissertation to my parents, Gregory L. Polyn and Patti L. Polyn.

Chapter 1

An overview of the thesis and the literature

1.1 Overview of the thesis

The conceptual focus of this thesis is the ability of humans to target episodic memories, i.e., to reaccess past states of the world encoded by the memory system. I describe a framework for understanding this process that relies on the interaction of three cognitive systems in the brain: semantic memory, episodic memory, and a context maintenance system that acts to probe the episodic memory system. I build upon previous studies of these systems, in which each of these cognitive systems is mapped onto a particular anatomical region of the brain. Respectively, these areas are posterior cortex, medial temporal lobe, and prefrontal cortex. This framework is investigated in a series of studies using the free recall paradigm. I present an imaging study that looks for evidence of memory targeting in the activity patterns of the brain; a behavioral study that attempts to manipulate the accessibility of memories by perturbing the context maintenance system; and a modeling study that attempts to codify this framework as a connectionist model.

In this first chapter I review relevant behavioral, modeling and neuroscientific studies of the episodic memory system, and the neural systems thought to comprise this cognitive system. This

review is followed by a consideration of the details and mechanisms of the current model in the context of this literature.

In the second chapter, I describe a neuroimaging experiment in which I use pattern classification methods to track the second-by-second fluctuations of patterns of brain activity during free recall. This study provides a methodological foundation for neuroimaging studies of contextual reinstatement processes during verbal recall. The results of this study highlight the idea that context effects on memory organization can be carried out by any brain area that has both a relatively stable pattern of activity during the encoding process, and connections with the medial temporal lobe brain regions.

The third chapter describes a behavioral study of the effect of encoding task on memory accessibility during free recall. Specifically, I explore the hypothesis that by changing encoding task midway through the study list, the sets of items encoded in the context of each task become isolated from each other in memory. The isolation of these memories from each other is quantified by calculating the probability of the subject making an output transition that crosses the mid-list task boundary. This probability is significantly reduced in the task-shift condition, relative to the control condition, in which all items are encoded in the context of the same task. This suggests that patterns of task-related activity act as an effective context, uniting memory traces encoded within-task and isolating traces encoded under different tasks.

The fourth chapter describes a connectionist model of the interactions of posterior cortex, prefrontal cortex and the medial temporal lobe during free recall. I describe a series of simulations that illustrate how the prefrontal system can manipulate the medial temporal memory system to flexibly extract previously encoded memories. The model also captures the "memory isolation" effects seen in the encoding task-shift experiment described in Chapter 3. Finally, I present simulations that attempt to capture the pattern of deficits seen in elderly subjects in free recall (described in Kahana et al., 2002). Damage to the gating system in the prefrontal component produces simulated behavior consistent with the observed pattern of deficits.

In the final chapter I describe in more detail the ways in which these studies inform each other, as well as a series of future experiments and simulations that investigate some of the core predictions of this framework.

1.2 Review of the literature

The proposed framework attempts to both describe the behavior of cognitive systems and relate these cognitive systems to particular brain areas. To this end, I review three regions of the literature. First, I review a series of behavioral studies of the free recall paradigm which have been used to guide construction of formal models of the memory system. Second, I review these models, and the mechanisms that have been implemented to describe human behavior in these tasks. Finally, I review a subset of the neuroscientific literature that speaks to the functional role of the prefrontal cortex in episodic memory.

A large literature has grown up around the idea that frontal areas are involved in the strategic manipulation of the memory system, to serve the behavioral goals of the organism (reviewed in Moscovitch & Winocur, 2002). These memory processes are described as "Working with memory", and a set of functions have been proposed for the various areas of frontal cortex. However, there has been relatively little work by way of formalizing the mechanisms responsible for carrying out these functions. I propose a simple iterative interaction between the frontal and medial temporal areas of the brain that gives rise to a rich set of memory-searching behaviors. Within the framework of Moscovitch and Winocur (2002), these processes may fall under the heading of cue maintenance and updating. However, the memory search mechanism proposed here may also explain behavioral variance usually ascribed to inhibitory processes (see section 1.2.1). Furthermore, I will review a theory by Norman and Schacter (1996) in which they propose that a contextual focusing mechanism, quite similar to the one described here, can also play the role of a memory monitoring system, another theoretical construct that is often considered distinct from a cue maintenance or contextual focusing system.

Before delving into the literature, it is worth considering the utility and even necessity of having a contextual focusing system to manipulate the episodic memory system. Every day, we store hundreds of new memories, in a variety of spatiotemporal contexts. Sometimes these memories can be retrieved and examined effortlessly, but sometimes, to our frustration, we find our efforts blocked, and our memories inaccessible. What are the mechanisms by which the brain pulls memories from seemingly limitless stores, and what factors contribute to their degradation and inaccessibility? Our theory posits that prefrontal cortex (PFC) stores a "targeting vector" or "context vector" that allows the memory system to focus on appropriate contexts. By way of establishing the necessity of a targeting mechanism, consider the scenario in which you must locate where you carelessly dropped your house keys the night before. You have many memory traces involving your keys, and a large number of these involve your keys in the environmental context of your house. In order to effectively retrieve the proper information, you must focus your memory on a specific temporal window; by thinking about the types of things you were doing last night (e.g. washing dishes), you can retrieve the appropriate episode (e.g. the keys are by the kitchen sink). The everyday strategy of mentally retracing one's steps allows one to reinstate specific internal contextual states, focusing memory search on the set of traces that occurred in a specific spatiotemporal context.

The concepts of context and cuing have played a major role in studies and theories of the human memory system. These concepts will be elaborated in great detail throughout this chapter. For our purposes, context is the set of features present along with a given to-be-remembered stimulus. As mentioned above, this can correspond to external features (the environmental context) as well as internal features (inner mental context). As such, context is considered as a *representation*, a set or vector of features active in the cognitive system (see section 1.2.2 for more detail). This is distinct from the representation of an item or stimulus, which is also considered as a vector of features concurrently active in the cognitive system. Given the characterization of context as a representation, cuing can be considered an *operation*, something that the cognitive system does to pull specific information out of the memory system. Section 1.3 describes a particular mechanistic account of a cuing operation, one in which the representation of context is used to cue for stored

item or stimulus representations. As will be reviewed below, one can draw conclusions about the structure of the memory system by observing the situations in which perturbation of context affects memory accessibility.

Given these definitions of context and cuing, I can sharpen the definitions of memory targeting and memory search. Memory targeting is the operation by which a set of stored memories can be made more accessible by reinstating in the cognitive system some representation present when the memories were originally encoded. Thus, in a memory experiment, reinstatement of the study context representation allows one to target the set of memories encoded in that context, by using that representation to cue the memory system. Finally, memory search is the series of cognitive operations by which the cognitive system pulls a set of memories from its stores. In the subsequent sections, these concepts will be considered in terms of the behavioral literature, the modeling literature, the neuroscientific literature, and the current model of episodic memory targeting and search.

1.2.1 Behavioral studies of memory targeting

In this section I review the ways in which these concepts of context, cue, targeting and search have been explored in the behavioral literature.

The major paradigms

I partition the space of memory paradigms into three sections (recognition, cued recall and free recall), based on how the memory system is probed by the experimenter during the recall period. It is worth setting up each paradigm before reviewing the broader literature, as each paradigm has been used to investigate the issues of context and cuing in memory.

Recognition. In a standard recognition paradigm, the recall period is broken into trials. Each trial contains the presentation of a stimulus, which must be judged as "new" (never before seen), or "old" (previously presented) in the context of the experiment. Contemporary theories of recognition memory explain behavioral performance in this paradigm in terms of two systems, a familiarity

system and a recollection system (Yonelinas, 2002; Norman & O'Reilly, 2003). These systems are thought to correspond respectively to medial temporal cortical areas and hippocampus proper (Aggleton & Brown, 1999; Norman & O'Reilly, 2003). By these theories, a given old/new response is made on the basis of information provided by both systems. By these theories, the two systems have radically different outputs. The familiarity system provides a scalar signal that can be interpreted as the likelihood that a given stimulus has been seen before, whereas the recollection system provides a reconstruction of the original representation of the stimulus. Furthermore, the output of the familiarity system is thought to be less sensitive to changes in context than the recollection system (Norman & O'Reilly, 2003; Polyn, Norman, & Cohen, 2002).

Cued recall. In a cued recall paradigm, the recall period is divided into trials. On each trial a stimulus is presented, however, the judgment is based on the recall of something associated with the stimulus. In the case of paired associate recall, subjects are asked to retrieve the identity of a stimulus that was paired with the test stimulus at study. In the case of source memory, the subject is asked to recall some component of the original context in which the test stimulus was studied. This original context is broadly defined, and can be related to the study task, stimulus modality or some environmental feature set present at study.

In this paradigm, the familiarity system is less able to contribute. The scalar signal provided by this system may signal to the subject whether she has seen the test item, but it will not provide information about other features associated with the test item at study. Thus, in this paradigm it is thought that the subject must rely on the output of the recall or episodic memory system (Polyn et al., 2002).

Free recall. Free recall differs from these other paradigms in a few key ways. First, the recall period is not divided into trials. The recall period is preceded by a vague cue (e.g. "Recall words from the most recent list"), and so, the contextual focusing system is forced to probe the memory system; the external environment provides minimal cues at best. Free recall thus provides a test of the cognitive system's ability to both create an appropriate cue, and use that cue to extract informa-

tion from the memory system. One can draw conclusions about the nature and fidelity of this cue by examining the probability of recalling certain list items, as well as the probability of transitioning between sets of list items during recall (see chapter 3).

Performance in a free recall paradigm is most often examined in terms of serial position effects, the probability that a subject will recall a study item, given the item's original position on the study list. The two major serial position effects are the recency effect, an enhanced probability of recall for the final items of the list, and the primacy effect, an enhanced probability of recall for the first few items of the list. The recency effect is easily disrupted, for example, by some kind of distraction interspersed between study and test (Glanzer & Cunitz, 1966). The conditions under which the recency effect is disrupted have had a major effect on theories of context and memory. These theories and the associated data are discussed later in this section.

Context change and cuing processes in free recall

The memory search process can be characterized as driven by a cue, used to probe the memory system. I now elaborate upon how the concepts of "memory cue" and "memory context" are intimately related.

For decades, researchers have described the nature of cuing processes in memory. The human memory system is often described as creating associations between stimuli; an episode is considered to be a bundle of associations among the various stimuli present at a moment, including environmental features, attended items, and states of mind (Tulving, 1983). A defining feature of an episodic memory is that, given the presence of a small subset of the original stimuli, the entire set of the associated stimuli can be brought back to mind (in the cognitive literature this process is referred to as redintegration, and in the connectionist literature it is referred to as pattern completion). Thus, when one returns to a long-ago frequented environment, such as a childhood home, one finds it crowded with usually dormant memories; the features of the environment (the environmental context) are acting to cue the episodes that were encoded within those walls. In this section, I explore the types of representations that can act as context, and the conditions under which contextual rep-

resentations are used as cues to the memory system. These issues are elaborated in terms of context change paradigms. There are a broad range of paradigms whose results can be interpreted in terms of theories of contextual change.

I first review three versions of the context change paradigm, in which researchers have manipulated environmental context, task context, and a more nebulous "inner mental" context. I then follow this with a context-based interpretation of the Brown-Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959), and a discussion of a classic memory experiment often taken as evidence for context-based memory targeting (Shiffrin, 1970).

Changes in environmental context have a large effect on memory accessibility. Consider an experiment in which a subject studies a set of items in testing room A, and then moves down the hall to study another set of items in testing room B; the subject, still in room B, is then asked to recall as many items as she can. There are two main effects of a context change of this nature. First, items studied in room A will have a reduced probability of recall. Second, items studied in room B will have an enhanced probability of recall, as the A-items are not interfering with retrieval (Godden & Baddeley, 1975; Smith, 1988). While clear context effects have been seen in tests involving free and cued recall (Smith, 1988; but see Fernandez & Glenberg, 1985), they have been more difficult to produce in tests of recognition memory (Smith, 1988; Murnane & Phelps, 1993, 1994, 1995). This may be due to a number of factors. First, as mentioned, the familiarity system may be less sensitive than the recollection system to changes in context between study and test. Furthermore, even if the recollection system is sensitive to changes in context, during a recognition test the original item is presented to the subject, giving the recollection system a chance to retrieve the associated context before the subject makes a decision.

Even given the sensitivity of subjects in a free recall paradigm to changes in environmental context, there are cases in which the context change effect is quite fragile. Bjork and Richardson-Klavehn (1989) define a set of paradigms where the environmental context is incidental to the study period. In their words, incidental context "does not influence the subject's interpretation of . . . the

target material at encoding". In these paradigms, context change effects on recall performance are inconsistently reported, and are often statistically weak (Bjork & Richardson-Klavehn, 1989). In terms of the current framework, the context representation will only affect memory accessibility to the extent it is incorporated into any memory traces that are laid down at study. If a context representation is incorporated into memory traces, then the presence of that representation at recall will enhance the probability of retrieving the set of memory traces encoded in that context (and make those traces less accessible in its absence).

In this dissertation, I present a memory system that is cued by active representations in the broader cognitive system. As such, there is no fundamental difference between representations of the external environment, and representations of the internal environment, that is, the collective states of all the non-environmental areas of the brain that project into the memory system. This inner mental context has had a place in theorizing about the memory system for many years (c.f. Bower, 1972). The history of its conceptual evolution will be considered in the next section.

Before I describe the effects of internal context change on memory accessibility, it is worth considering the expected differences between external and internal context. Certainly the external environment contains a major unchanging component (to the extent that a subject stays put during an experiment), however, there will also be components that change due to the passage of time and a drifting of attention (Estes, 1955). We have similar expectations of internal context. Theories of task performance posit that stable patterns of activity in prefrontal cortex are used to guide the processing of stimuli in task-appropriate ways (in paradigms such as Stroop: Cohen, Dunbar, & McClelland, 1990; and AX-CPT: Frank, Loughry, & O'Reilly, 2001). Furthermore, in cases where the task involves processing trial-unique items, there will also be components of the context that evolve over time.

It seems that in a memory experiment, one of the most powerful components of internal context will be related to the encoding task. For example, consider the study of Watkins and Peynircioğlu (1983), in which items encoded with three distinct tasks were interspersed in a study list. Following

the study list, there were three successive recall periods, in which items were recalled by task. Each of the three recall periods showed a distinct recency effect for the studied items, despite the fact that for two of these recall periods, the subject was quite distracted by the process of recalling items from another task. This distraction was expected to have eliminated the recency effect (Glanzer & Cunitz, 1966). Thus, it seems that task representations can provide a strong cue for items. These results will be revisited below, in the discussion of serial position effects and theories of context. Furthermore, in chapter 3, I describe a behavioral experiment in which subjects study items in the context of two encoding tasks in a free recall experiment. I show that the nature of the encoding task has a strong effect on both the accessibility of the items and the order in which they are recalled.

Above, I introduced the concept of inner mental context as the set of representations active in the cognitive system that do not directly reflect the features of the external environment. Above, I propose that task representations form a major part of inner mental context. However, it seems that changes to other components of inner mental context can also influence memory accessibility, as will be elaborated below.

The final set of experiments described in this section are variants of the directed forgetting paradigm (MacLeod, 1998). In this paradigm subjects study two sets of items. During a break between the two sets, subjects are told that the first set of items were practice, and can be forgotten; this is referred to as the forget cue. In the control condition subjects are either told nothing during the break, or are reminded to remember all items. Regardless of experimental condition, subjects are asked to recall items from both sets during the recall period that follows. Interestingly, in the case where subjects are told to forget the first set of items, they have poorer memory for these items during the recall period (Geiselman, Bjork, & Fishman, 1983). It appears that a brief verbal instruction is enough to cause forgetting of a set of stored memory traces. Furthermore, when told to forget the first set, the subjects have better memory for the second set of remaining items, as if the forget items are not contributing to interference at test (Geiselman et al., 1983). These two effects are known as the costs and benefits of directed forgetting (MacLeod, 1998).

In a review of directed forgetting, Bjork (1989) posits a retrieval inhibition mechanism as an explanation of the phenomenon. By this account, memory traces of the first set of items are somehow suppressed in the forget condition. Bjork (1989) is quite explicit that inhibition is meant in a strong sense, referring to suppressive brain activity. However, a recent paper by Sahakyan and Kelley (2002) provides an explanation based on context change. By their account, the forget cue causes subjects to attempt to think of something unrelated to the current experiment (such as a recent wedding), causing their inner mental context to be changed. Thus, the second set of items are encoded in a new inner mental context, causing the two sets of items to be isolated from each other in memory. In order to investigate this hypothesis, Sahakyan and Kelley (2002) used a variant of the directed forgetting paradigm. Instead of asking subjects to forget the first set of items, they asked subjects to perform elaborative mental activity (such as imagining what they would do if they were invisible). This mental activity produced results resembling both the costs and benefits of directed forgetting (decreased recall of the first set, relative to the control condition, and increased recall of the second set, relative to the control condition).

It is interesting that mental activity between the study lists would reduce the accessibility of items encoded prior to the mental activity. As will be described in sections 1.2.2 and 1.3, a number of models of human memory posit that context representations which are present in the cognitive system during encoding of a set of memory traces can be used to retrieve those memory traces during a later recall attempt. By this theory, the mental activity between the study lists in the Sahakyan and Kelley (2002) paradigm would perturb the context representation, such that the context representation active during the later recall period would be less consistent with the memory traces corresponding to the first set of items.

In the directed forgetting paradigm there are two other data-points that support the idea that the main effects are produced by some sort of change to inner mental context. First, the effects of the forget cue are reduced if items from the first set are presented at test (MacLeod, 1998). In this case, it is possible that upon presentation of these first set items, the recollection system retrieves context

that was associated with them. This retrieved context can be used to drive further recall items from this first set. Second, if a forget cue is given, but no second set items are presented, there is no decrement in recall for these first set items (no cost of directed forgetting; Gelfand & Bjork, 1985). This suggests that the first set context is not truly "lost" (or overwritten) until some second set items are presented. This context based account of directed forgetting will be briefly revisited in chapter 5, where it can be fully elaborated in the context of the current model.

In all of these paradigms, a context-based theory provides a good account of the observed phenomena. This is most remarkable in the case of the directed forgetting paradigm, in which an inhibition-based account seemed the most reasonable for many years (Bjork, 1989). In all of these cases, the behavioral phenomena can be explained in terms of the degree to which contextual representations allow the subject to focus on particular sets of stored memories. This ability to focus on certain sets of memories to the exclusion of others may be used to explain proactive interference effects in memory.

Proactive interference describes the phenomenon in which the encoding of a set of memories reduces the probability of recalling a later encoded, but similar memory (Greene, 1992). The behavioral results associated with the Brown-Peterson paradigm are often explained in terms of proactive interference (Wickens, 1972). This paradigm is quite similar to the free recall paradigm. Subjects study a short list of items (often between 3 and 5), drawn from a pool of similar items (examples of item classes: consonants, digits, words). There is a distraction period after the study period, in which the subject performs an unrelated task (such as counting backwards) designed to keep recall performance off ceiling. This distraction period is followed by a recall period, in which subjects are asked to report items from the most recent list. It was found that as subjects studied more and more lists with items drawn from the same class, recall performance declined. However, if the item class was changed, recall performance recovered. This recovery was called "release from proactive interference" (Wickens, Born, & Allen, 1963). Recently, theorists have suggested explanations of this pattern of results based on temporal discriminability (Baddeley, 1990; Greene, 1992). By these

theories, when a subject is asked to recall the same class of items again and again, it becomes difficult for them to discriminate between the most recently presented set and other recent sets. When the item class is changed, the most recent set is of a different type from other recent sets, and the recall process is much easier. The contextual focusing mechanism described in section 1.3 provides a good explanation for these findings, if one posits that item class becomes part of the context used to probe memory. As we shall see in the description of the TCM model in section 1.2.2, there is reason to believe that item-related information is integrated with context.

As mentioned, in all of these paradigms, context is being used to focus on a particular set of items in memory. In the case of the Brown-Peterson paradigm, this can prove detrimental to performance, when the items are all drawn from the same class. It is worth considering that a standard multi-list free recall paradigm is not unlike the Brown-Peterson paradigm. Each list is quite similar to the last, in terms of the task being performed, and the nature of the stimuli being encoded. In this paradigm as well, subjects often have difficulty focusing on the previous list. This difficulty is often exhibited as prior-list intrusions (Kahana, Dolan, Sauder, & Wingfield, 2005).

A classic study of memory targeting in multi-list free recall (Shiffrin, 1970) asked subjects to recall from the list-before-last. By varying the length of the lists, Shiffrin was able to show that the recall performance depended only on the length of the to-be-recalled list, and not the intervening list. This was taken as evidence that subjects could effectively focus their memory search on the list-before-last, to the extent that there did not seem to be any interference from the items on the most recent list. However, Ward and Tan (2004) replicated this paradigm using an overt rehearsal procedure, in which subjects are encouraged to say aloud any previous items that come to mind during the study period. They showed that subjects find this task extremely difficult, and tend to rehearse a few items from the list-before-last throughout the presentation of the intervening list. It seems possible that in order to properly focus on sets of previous memories to the exclusion of others, the sets of memories must have distinct contexts associated with them. In the case of the Shiffrin (1970) paradigm, where the items and lists are all quite similar, it may be that the contextual

focusing system is unable to target the list-before-last, and subjects settle on an alternate strategy to satisfy the experimental demands.

A contextual account of serial position effects

In the previous section I reviewed a series of paradigms in which changes in context affected the accessibility of memories, measured in terms of probability of recall of an item from a given set. Four major types of context were mentioned: environmental context, encoding task context, inner mental context, and item-associated context. In this section, I introduce a variant of encoding task context, the context associated with a distractor task. A distractor task (mentioned above in relation to the Brown-Peterson paradigm), is usually an effortful task, such as counting backwards by sevens from a three-digit number. The effect of distractor task on the recall of items from the end of a study list (the recency effect) has had a major role in shaping theories of context in free recall. These theories, and the associated behavioral findings, are reviewed below.

An examination of the serial position curve (Baddeley, 1998; Parkin, 1993) in immediate free recall yields three features: a flat mid-portion of the curve, and an enhanced probability of recall for both initial items (the primacy effect) and final items (the recency effect).

Studies using the overt rehearsal method, in which subjects are encouraged to verbalize any previous items that come to mind during study, suggest that the primacy effect is explainable by a tendency for initial items to continue to receive rehearsal throughout the list (Fischler, Rundus, & Atkinson, 1970; Ward, 2002). This is supported by the finding that primacy effects are completely removed in an incidental learning paradigm, in which subjects were unaware of an upcoming memory test and thus unmotivated to rehearse (Glenberg et al., 1980). As such, the primacy effect has played a relatively minor role in shaping theories of context and memory. However, it is interesting to note that rehearsal processes can be thought of as brief bursts of recall undertaken by the subject during the study period, and therefore susceptible to the same context effects as reviewed elsewhere in this section.

Early theoretical accounts of the recency effect in free recall posited two memory systems, a

short-term and a long-term system (Waugh & Norman, 1965; Glanzer & Cunitz, 1966). By this view, items from the end of the list were primarily recalled due to the operation of this short-term mechanism. This short-term store was characterized as a sort of buffer: it had a relatively small capacity, a high probability of recall for items stored in it, and was easily disruptable. This characterization was supported by findings from the delayed free recall paradigm, where subjects were asked to perform a distractor task between study of the items, and the recall period (Glanzer & Cunitz, 1966). Serial position curves derived from this paradigm showed a severely attenuated recency effect; this was interpreted as being due to the disruption of this short-term store. [1]

Bjork and Whitten (1974) created a variant of the basic free recall paradigm called continuous distractor free recall. By instructing subjects to perform a distraction task after each list item, they hoped to both disrupt the ability of the subjects to rehearse, and wipe clean the short-term store, resulting in a completely flat serial position curve. However, the resulting serial position curve was anything but flat; in fact, it retained all of its major features, including both primacy and recency. Bjork and Whitten (1974) ended up with the same curve, but now with no theoretical mechanism to explain the data.

A number of context-based accounts sprang up to fill the conceptual void. The first of these attempted to explain the observed serial position effects in terms of the passage of time (Glenberg, Bradley, Stevenson, Kraus, Tkachuk, Gretz, Fish, & Turpin, 1980; Glenberg & Swanson, 1986). Specifically, these accounts attempt to explain recency effects as a function of the ratio of time between the presentation of successive items (the inter-presentation interval or IPI), and the amount of time between the end of the study list and test (the retention interval or RI). By these accounts, items become less accessible as more time passes since their study. The amount or type of distraction that fills the time is not the relevant variable, but simply the amount of time that passes; distraction is

[1] Much of the debate of the time was over whether this short-term store was a necessary theoretical construct, or whether context-based accounts could completely handle all existing data. There is some evidence for a buffer-like component in the cognitive system (Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005). However, as I am considering a set of paradigms designed to minimize or eliminate the involvement of a short-term store, it shall not be considered further in the current write-up.

just a convenient way to keep the subject from rehearsing. There is value to this depiction, however, it is not the entire story, as elaborated below.

Even with both IPI and RI held constant, Koppenaal and Glanzer (1990) showed that one could significantly reduce the size of the recency effect by changing the nature of the distractor task between study and test. By their account, all continuous distractor free recall studies had used well-practiced distractor tasks; subjects were able to get used to this task, and still employ a buffer-like process despite the distracting activity (Koppenaal & Glanzer, 1990). By switching to a different distractor task at the end of the list, they disrupted the buffer and reduced recency.

This interpretation was called into question by a subsequent study by Thapar and Greene (1993), in which subjects performed distractor tasks that they had not practiced at all. The standard context-based recency effect was shown. They replicated the finding that the item preceding a distractor task switch was less well recalled, and also showed that the distractor task switch could occur mid-list, to much the same effect. They went on to show that even with a different distractor task following every item in the list, there was still a recency effect. While I will present a task/context based explanation of these findings, it is worth noting that Thapar and Greene (1993) had a different explanation. They posited that the switch in distractor task confused an unspecified context system, causing it to link the distractor item and the previous study item to the same contextual state; this contextual cue was thus overloaded, and the item was less likely to be recalled. I argue that an alternate context-based explanation is more appealing: the new distractor task causes a shift in inner mental context, and this new mental context is a relatively poor cue for items encoded before the switch.

A study by Petrusic and Jamieson (1978) provides some interesting data for this account. They ran subjects in a standard delayed free recall paradigm, with a distractor task interpolated between study of the list and recall. They showed that as the difficulty of this interpolated task increased (from listening to music to shadowing auditorially presented digits), the recency effect showed a continuous decline. By our account, Petrusic and Jamieson (1978) were varying the degree to which

inner mental context was being disrupted, and therefore the degree to which it was a good memory cue. In fact, they were able to show that the most demanding distractor task (digit shadowing) reduced recall at every serial position in the list.

In order for a representation in the cognitive system to play the role of context, it needs to have some relatively stable components across the set of encodings created during study, and it needs to be present at recall. In the above account, the distractor task plays two roles. In the cases in which there is a single distractor period between study and test, it seems to disrupt context, and reduce recall of the list items (Glanzer & Cunitz, 1966; Petrusic & Jamieson, 1978). However, when the distractor task occurs after every item on the list, the distractor task becomes part of the context of the study items. In this case the distractor task shows no disruptive effect; the recency effect is intact (Bjork & Whitten, 1974). The distractor task, originally introduced to segregate the items of the list, and prevent rehearsal, can actually serve to bind them together in a common context. This theory posits not only that distractor task information forms part of inner mental context, but also that its effect on context persists beyond the moment of task performance. The mechanisms by which a representation can persist in the cognitive system are reviewed in section 1.3, and the issue of distractor task effects on memory accessibility is revisited in chapter 5.

A contextual account of output transitions in free recall

There is one more class of free recall phenomena that I would like to cover before turning to the review of models of context in memory, which is that of output transitions during recall. The studies reviewed above focused mostly on serial position effects, that is, the probability of recalling an item given its position in the study list. However, this measure of recall probability obscures the behavioral dynamics that occur during the recall period. Studies of the series of responses made during the recall period have shown that the probability of reporting any one item is affected by the identity of other recently reported items (Kahana, 1996; Howard & Kahana, 1999). The major phenomenon is the lag-recency effect (Kahana, 1996); subsequent recalls tend to come from nearby serial positions to previous recalls. This process is measured with the conditional response

probability curve (CRP); this curve displays the probability that a subject, having just recalled item N from a list, will next recall item N+1, N-1, etc. This phenomenon receives greater attention in chapter 4 (see figure 1.2 for an example of a CRP curve).

The lag-recency effect shows that the probability of recall for items in the list changes during the recall process. This implies that during the recall process, the cue being used to retrieve memories is being altered. Howard and Kahana (1999, 2002a) develop this idea in their theory of Temporal Context Memory (TCM). The mechanism underlying TCM will be elaborated in the next section (1.2.2). However, I introduce the main features of the theory here, in order to better explain the behavioral data.

TCM describes a simple memory system containing item representations and context representations. During the study period, an item representation is presented, and two things happen. First, the context representation is altered slightly, due to the presentation of the item. Second, the item representation is bound to the current context representation. Thus, the context representation changes slowly over the course of the study period, and at each iteration is associated with another of the study items (Howard & Kahana, 2002a).

TCM suggests that during the recall period, there is a bidirectional interaction between the memory system and the context system. At the beginning of the recall period, the end-of-list context state is used to retrieve an item representation. Given retrieval of an item representation, the context representation is updated. If a subject recalls item N, the memory system retrieves the context representation associated with this item. Since the context representation changes slowly over the course of the study period, this updated context representation is a good cue for items studied nearby in time to the just-recalled item. This allows TCM to capture the lag-recency effect (Howard & Kahana, 2002a). Howard and Kahana (2002a) describe a mathematical model that implements TCM and produces simulated data that is an excellent fit to behavioral observations. This model builds upon a long tradition of mathematical models of context in memory, as will be reviewed in the next section.

The major points of this section can be summarized as follows: Memory retrieval is a probabilistic process, and the probability of retrieving a particular memory varies with its accessibility. During free recall, retrieval is driven by a context-based cue. The accessibility of a given memory varies with its congruence with the current cue. In the case of the recency effect, the end-of-list context cue is more congruent with the most recent items than with more distant items. Given a distraction interval, the context cue is disrupted, and recent items lose their special advantage. However, if the distractor task is performed throughout the list, it becomes part of the list context and the recency advantage is spared. Finally, the context cue is updated during the recall period, which explains the tendency of subjects to successively report items studied nearby in time.

1.2.2 Context in models of human memory

The previous section raised several theoretical issues in the context of the behavioral literature and proposed how a memory model might explain these findings, specifically, by positing a context-based cuing mechanism. Context-based accounts have been used by a number of researchers to explain puzzling phenomena. I now review some of the existing models that have used context mechanisms to explain the phenomena of memory search. Contextual mechanisms in memory models have been handed forward through the literature for many decades, with each researcher tweaking the implementation to fit a new domain of data. In the previous section I referred to context and item representations and the ways in which they are manipulated by the cognitive system. A major contribution of the early mechanistic theories of human memory was to formalize the definition of a representation as a vector of features, in which each feature is represented by a scalar or binary value (see, e.g. Bower, 1967).

Estes (1955) described how a time-varying context vector can be used to explain several puzzling phenomena of memory retrieval. The vector is composed of a number of binary elements whose activation state is determined by a set of probabilities. This theory was presented in the context of classical conditioning, and presented the vector as a set of stimulus-related features that would be associated with responses. Thus, his formalization is referred to in the literature as “stimulus

fluctuation theory". In this formalization, context varied only with the passage of time. This idea appeared elsewhere in the literature as well. Yntema and Trask (1963) proposed that encoded items might have "time tags" appended to them, to facilitate time-based search and recency judgements (however, they did not implement their hypotheses with a formal model).

Bower (1972) presented a model based on stimulus fluctuation theory and showed that it could be applied to the data of recognition memory paradigms, to determine whether an item came from a particular list. In the same paper, he applied the model to the recency data of Yntema and Trask (1963), and showed a good fit. These studies were among the first to explore the idea that memory traces could be described as sets (or vectors) of features.

In this work, Bower (1972) describes a general encoding process operating on an environmental stimulus. In his depiction, an internal representation is generated, which can be described as a vector of stimulus elements. The state of this representation can be influenced by, among other things, incidental features of the environment as well as the subject's inner mental state (Bower, 1972). Two corollaries follow: First, the same item, studied under two different states of inner mental context, will be encoded differently in the memory system in each case. Second, different items studied under the same state of internal context will be encoded similarly in the memory system. Many of the principles developed in these works are fundamental to modern connectionist investigations.

Raaijmakers and Shiffrin (1981) presented the Search of Associative Memory (SAM) model. This model has been used to fit much data from the domain of free recall. The goal of this project was to specify the retrieval and search algorithm, rather than the mechanism of storage. As such, the storage mechanism is implemented as a matrix of probabilities (the strength matrix). Given a certain cue or set of cues, the strength matrix determines the set of probabilities of retrieving each item from the list. While the SAM theory contains a cue representing general context, the original implementation did not explore fluctuations of this contextual element vector over time.

Mensink and Raaijmakers (1988) modified the SAM implementation by adding a time-varying

context vector, comprised of a set of binary features. This model was used to explain a number of findings in the interference and forgetting literature, including those from a number of cued recall studies of the AB-AC variety. In this implementation, the context vector was static within a given list, as the authors were not interested in accounting for within-list effects in that study.

The theory of Temporal Context Memory (introduced in section 1.2.1) introduced the notion of a context vector that changes over the course of the study list (Howard & Kahana, 2002a; Howard, Fotedar, Datey, & Hasselmo, 2005). This evolution is driven by features associated with the study items. This model is used to explain the detailed pattern of output transitions during free recall. I will now elaborate the mechanism of this model in more detail.

According to the TCM model, the drift of context over the course of the study list is driven not by random fluctuations, but by prior contextual representations associated with the items (an item's pre-experimental context). During study, this item-specific pre-experimental context is added to the current context representation, causing a small perturbation in the overall context representation. When an item is studied, the item representation is also associated with the currently active context representation (the experimental context).

When an item is retrieved during the recall period, the context representation is updated by vectors arising from two sources, the pre-experimental context and the experimental context. By scaling the relative influences of these two vectors during recall, Howard and Kahana (2002a) were able to explain the tendency for subjects to make output transitions in the forward direction. This bias to make forward transitions is a constraining data-point and deserves special attention. Given recall of the item from serial position N in the study list, a subject is more likely to next recall item N+1 than N-1. As mentioned above, the TCM model has two factors driving context update at recall. Reinstatement of the pre-experimental context biases the system to make forward transitions. This is because during study, this pre-experimental context is integrated with the existing vector and then becomes associated with all successive items. By reinstating the pre-existing context at recall, it biases the system to next recall an item from one of these successive positions. In contrast,

the retrieved experimental context has features that are associated with previous items as well as features associated with successive items. When the context representation is updated by both of these sources, the net effect is a bias for the system to make forward transitions. This explanation will be returned to in chapter 4.

Becker and Lim (2003) present a connectionist model examining the role of prefrontal cortex (PFC) during free recall. This model investigates the role of semantic category information in memory storage and search by modeling subject performance in the California Verbal Learning Task (CVLT). The model contains a PFC component that learns to develop representations that are used to cue memory for item representations drawn from a particular semantic category. As mentioned above in section 1.2.1, any type of representation can be used to cue the memory system, as long as the representation has a stable representation across sets of items. Thus, environmental context, task context, temporal context, and semantic context may be used in similar ways by the cognitive system. I will return to the Becker and Lim (2003) model in the general discussion in chapter 5.

1.2.3 Neuropsychological studies of memory targeting

As mentioned, the current model attempts to make contact between the cognitive systems and the anatomical systems of the human brain. In this section, I review some of the relevant neuropsychological studies of frontal involvement in memory search and retrieval.

The difficulties of frontal patients on tests of memory have been described as an inability to properly focus on specific spatiotemporal contexts (Schacter, 1987) and semantic contexts. Frontally damaged patients are able to perform near the level of controls on tests of recognition memory, but when the memory tests require specific associated contextual information (source memory and cued recall), or require a guided search through memory (free recall), deficits are seen (Wheeler, Stuss, & Tulving, 1995). Despite the heterogeneity of frontal deficits, patterns of impairment begin to emerge. A note of caution is necessary. It is clear that different areas of PFC subserve different functions, but those functions are not yet precisely characterized. A number of lesion studies, by

necessity, lump together patients with different pathologies and different lesion sites. The degree of individual variation in deficits is often obscured by averaging procedures.

Source memory. Source memory is a form of cued recall, where contextual elements related to an item must be retrieved upon presentation of that item. During tests of recognition memory a response can be based on the familiarity of an item. During source memory a response must be based on monitoring of the contextual tag retrieved along with an item. This operation is presumably more demanding than those associated with recognition, and as mentioned above, could rely on the same PFC information involved in targeting. Two classic studies examine source memory in amnesic patients. A majority of the patients examined in these studies have conditions related to frontal impairment (Korsakoff's syndrome, anterior communicating artery rupture), but not all, so again, the results must be viewed with caution. Schacter, Harbluk, and McLachlan (1984) presented subjects with a number of facts about celebrities (e.g. Jane Fonda refuses to eat chicken), and after a delay tested their memory for both the facts and, for the correct responses, the source of their knowledge. The patients would often be able to recall the appropriate fact when asked (e.g. What does Jane Fonda refuse to eat?) but would fail to attribute the acquisition of the knowledge to within the experimental session. A similar study by Shimamura and Squire (1991) examined the performance of amnesic patients. Subjects were quizzed on obscure facts, and later tested for their retention of the knowledge. In the control subjects, there was a significant correlation between fact recall and source recall. However, in the amnesic patients, this correlation was well below significance, suggesting that the two processes (fact memory and source memory) can be dissociated. Some of these effects may be attributable to a deficit in monitoring of retrieved information, rather than an inability to access that information in the first place. These are both functions purported to reside within the frontal lobes (Moscovitch & Winocur, 2002). The mechanism of their interaction is as of yet unclear.

Cued recall. Frontal patients are often able to perform close to the level of control subjects on tests of cued recall, however, they show an increased susceptibility to proactive interference.

Shimamura, Jurica, Mangels, Gershberg, and Knight (1995) used an AB-AC paired associates test of memory, where a given item (e.g. queen) is paired with multiple associates over the course of the experiment (e.g. queen-king; later queen-crown). At test, the subject is required to report the most recent associate of a given item. Frontal patients are significantly impaired compared to control subjects on items that have multiple associates. A context-based account would suggest that this deficit arises from an inability to properly target the most recent memory (AC); competing representations (AB) are able to intrude and often are reported.

Free recall. Free recall is one of the most difficult memory tests, in that hippocampus has no good environmental cues with which to prompt memory retrieval. All cues must be internally generated. Frontally damaged patients show an interesting set of impairments on tests of free recall (Gershberg & Shimamura, 1995) including reduced overall performance, reduced category clustering (the tendency to report items from the same category successively), and reduced subjective organization (the tendency to recall words in the same order on multiple recall tests for the same list). In general they tended to use fewer organizational strategies to aid memory. The reduced overall performance is generally consistent with an inability to focus on the appropriate memories, but the reduced category clustering is more intriguing. [2] Interestingly, frontal patients showed spared serial organization, which is a tendency to recall items in the same order that they were studied. It is possible that this organization is mediated by the hippocampal system, which may still be able to form associations between consecutively presented items (as the paradigm they used did not prevent rehearsal). In this situation, a recalled word can act as a bottom-up cue and bias the system to retrieve words associated with it. Strategy instructions benefited the patients when given at either study or test, suggesting that the frontal patients were impaired at self-initiated application of strategies at both encoding and retrieval. Given that strategy instructions can cause a boost in performance in frontal patients, this also suggests that some frontal mechanisms may be present and

[2] As hypothesized above, PFC may be able to maintain activity in a particular node of a semantic network corresponding to a category common to a subset of the studied words, thereby driving recall of all items within that category (Becker & Lim, 2003).

functional, but not used appropriately by the system, unprompted.

Before completing this section, I will touch on an interesting disorder seen in the everyday behavior of a subset of patients with frontal damage – confabulation. Spontaneously confabulating patients (as defined by Schnider, 2001) will retrieve irrelevant memories and then act upon them as if they correspond to ongoing reality. It has been theorized that both aspects of this disorder can be attributed to a faulty "focusing" mechanism, which may be closely related to our targeting vector (Norman & Schacter, 1996; Schacter, Norman, & Koutstaal, 1998). In this view, an erratic targeting vector could cause random irrelevant memories to be retrieved from memory. Confabulation seems to be most common in patients with damage to orbitofrontal cortex (Schnider, 2001).

In the final part of this section, I review two studies attempting to characterize the performance of young and older subjects on common psychological tasks. While healthy aging is not technically a neuropsychological disturbance, the findings reviewed here do point to the possibility of certain deficits in the aged cognitive system.

The first study examines the performance of young and older subjects on the AX-CPT, a continuous performance task in which subjects must maintain and update a set of rules in order to respond correctly to stimuli presented on a screen (Braver, Barch, Keys, Carter, Cohen, Kaye, Janowsky, Taylor, Yesavage, & Mumenthaler, 2001). They posit that a wide array of cognitive deficits in the elderly may be related to a dysfunctional dopamine-mediated gating system in prefrontal cortex. The model used to simulate the task performance data in this study is a variant of the one presented in Frank et al. (2001) and used in the present simulations (discussed further in section 1.3). Specifically, Braver et al. (2001) suggest that elderly subjects tend to have deficits in both the representation and maintenance of context information over time. This context information is maintained in order to bias the cognitive system to respond appropriately to upcoming stimuli.

The second study examines the performance of young and older subjects in the free recall paradigm (Kahana et al., 2002; Howard, Kahana, & Wingfield, submitted). It is shown that older subjects have a decreased probability of recalling items at all serial positions in the list, relative to

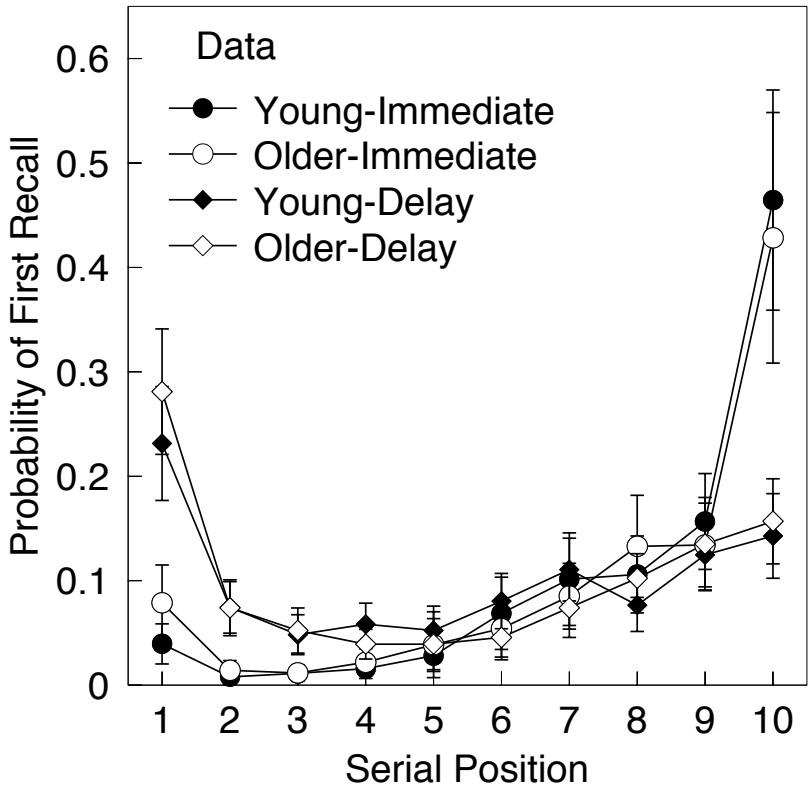


Figure 1.1: The probability of first recall curve, comparison of young to older subjects. (Figure provided by Marc Howard)

the younger subjects. Furthermore, older subjects do not show a significant bias to make forward transitions during recall (see section 1.2.1). This tendency is depicted in Figure 1.2. Finally, the older subjects initiate recall in the same way as the young. The probability of first recall (PFR) measure examines the probability that an item from a given serial position is the first item recalled. As shown in Figure 1.1, the PFR curves generated from the young and older subject groups do not show a significant difference.

Howard et al. (submitted) use the TCM model to fit this pattern of deficits. As reviewed in section 1.2.2, the context representation in TCM is updated using pre-experimental and experimental context. By reducing the influence of retrieved experimental context at recall, Howard et al. (submitted) are able to obtain a good fit to the elderly data. By the most recent account of TCM (Howard

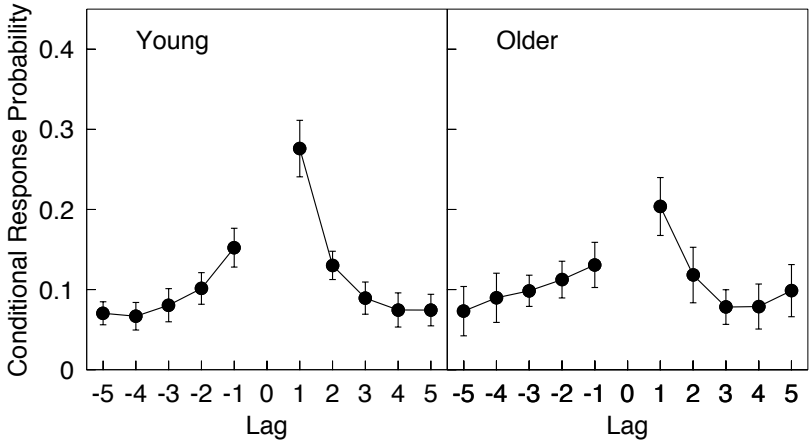


Figure 1.2: The conditional response probability curve in immediate free recall, comparison of young to older subjects. (Figure provided by Marc Howard)

et al., 2005), this component of context is retrieved by the hippocampus. Thus, they attribute the decline in performance in the older subjects to disrupted hippocampal function.

In chapter 4, I describe a series of simulations that attempts to fit this same pattern of deficits with the current model. Interestingly, damage to the hippocampal component of the current model does not capture the pattern of deficits seen in the older subjects. This pattern of spared PFR and flattened CRP is best captured by damage to the gating system of the PFC component of the model. The difference in conclusions between this account and the Howard et al. (submitted) account is attributable to structural differences between the two models. In the TCM model, the hippocampal component only serves to retrieve the experimental context representation, whereas in the current model, the hippocampal component retrieves item and context representations concurrently. Thus, damage to the hippocampal component of the current model hurts item recall, which flattens the PFR curve (see Figure 4.4(a)).

As will be elaborated in chapter 4, by damaging the gating system in the current model, the proper pattern of deficits emerges. Interestingly, the type of damage necessary to capture the elderly deficit in free recall in the current model is quite similar in character to the damage implemented by Braver et al. (2001), to capture the elderly deficit in task performance.

In summary, there is some convergent evidence for the idea that frontal dysfunction can produce a pattern of deficits consistent with a damaged targeting vector. However, as mentioned above, many of these studies of lesioned patients deal with quite heterogenous populations; it is unclear whether the same functional areas are damaged across patients. One of the benefits of the current approach is that one can begin to better characterize the various functions ascribed to PFC, leading to the possibility of devising tests that will dissociate these various functions. For example, Norman and Schacter (1996) suggest that a memory monitoring system and a contextual retrieval system might use the same machinery to achieve their goals, as a monitoring process most likely needs contextual information to assess whether a given retrieved memory is relevant to the current situation. In chapter 5, I describe how the current framework could be elaborated to implement a memory monitoring

system.

1.2.4 Neuroimaging of memory targeting

A number of researchers use neuroimaging techniques to analyze the contribution of prefrontal cortex (PFC) to episodic memory retrieval. I begin with a very brief description of fMRI methods, then I describe a set of criteria for considering an area related to targeting, finally I review relevant findings in the neuroimaging literature, and describe how distributed pattern classification will help us link these imaging studies to our theories of memory search.

Fundamentals of fMRI

Functional magnetic resonance imaging allows the detection of subtle changes in the oxygenation of blood across the subject's brain. These changes in oxygenation have been shown to be a good proxy for local field potential in an area (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001; Logothetis, 2003), which is in turn a good measure of processing in the local neural tissue. The scanner takes readings from 2-d slices of the brain; in a short time (about 2 seconds) it can take readings from slices covering the entire brain. In this case, 2 seconds is the time to repetition (TR) – the time it takes for the scanner to take another reading from the same slice of brain. Neural events are remarkably fast (on the order of milliseconds), so it could be asked what sense one could make of signals that are taken on the order of seconds. Thankfully, the MRI machine is not imaging the actual dynamics of neurons, but rather the hemodynamic response of the brain's vasculature to the activity of neurons in an area (which floods the local area with oxygenated blood). This hemodynamic response has a considerable time lag compared to the neural dynamics; some estimate the peak of hemodynamic response at 6-8 seconds after the neural activity that gave rise to it. While the temporal resolution is admittedly limited, fMRI gives quite good spatial resolution. The signal from the scanner can be localized on a millimeter scale. A standard voxel [3] size is 3mm by 3mm by 5mm. Finally, a major advantage of fMRI is the sheer number of simultaneous recordings; the

[3]The term voxel is short for volume-element, much like pixel, which is a picture-element in a 2-d image.

scanner can record signal from 30-40,000 distinct points distributed across the brain.

Neuroimaging human memory

In this section, I review a set of neuroimaging (specifically fMRI) studies of human memory in which components of brain activity recorded during the study period were shown to be reinstated during the recognition or recall period. For example, Nyberg, Habib, and Tulving (2000) and Wheeler, Petersen, and Buckner (2000) showed that during a memory task, cortical areas related to the modality of an associate to a word reactivate when a subject is presented that word in a subsequent recognition memory or source memory test. In other words, cortical activity seen during encoding is recapitulated during recollection. One important difference to note between these studies and the current one is that since the modality of study item was varied (visual versus auditory in Wheeler et al., 2000), large swaths of cortex could be counted on to differ between the items. In the current study, all items are presented in the visual modality; thus classification is based on more subtle differences between cortical states. It is important to note that other researchers have tread this ground as well. Kahn, Davachi, and Wagner (2004) show, in a source memory paradigm, retrieval activity related to encoding task.

In the previous sections, I defined a context representation as one that is present during the study period, and whose presence at test increases the accessibility of congruent memory traces. I also defined memory targeting as an operation by which a context representation is used to probe the memory system. To argue that a pattern of brain activity is being used for memory targeting, it is important to correlate some measure of recall performance with the strength or presence of that pattern.

It is difficult to find evidence for or against the existence of an area that meets these criteria by reviewing the neuroimaging literature. Often, published studies do not report activity of an area at encoding, or only areas that are differentially activated by encoding and retrieval are reported. However, there are a few notable examples of studies where activity patterns show a profile that is generally consistent with a memory targeting process. Dobbins, Foley, Wagner, and Schacter

(2002) reported an area in left anterior ventrolateral PFC that activated over a baseline during both a semantic encoding period and a source recollection period. Henson, Shallice, and Dolan (1999) reported a right dorsolateral PFC area whose activity increased over baseline during the encoding task, increased further still during an item recognition period, and increased even more during a source recollection period. Unfortunately, neither of these studies correlated recall or recognition performance with the moment-by-moment strength of the potential targeting-related activity.

A final note regarding the fact that these two studies found different areas of PFC to show this recall-related response profile. These two studies used different encoding tasks (semantic judgement vs. location on screen). Thus, I do not expect that the same area of PFC would be reported in each study. The point of the previous sections was to establish the idea that representations of many types can be used to target memories.

There are very few published studies of free recall in the scanner. The few imaging studies that investigate free recall scan during encoding, but test subjects outside the scanner, making it impossible to investigate the reinstatement of encoding related representations. In chapter 2 I describe an fMRI study of free recall in which distributed patterns of brain activity detected during the study period are shown to be reactivated at test and correlate with the recall behavior of the subject. The activation of these patterns seems to precede recall verbalization by a few seconds, strengthening the idea that these patterns are related to memory search (see also Polyn, Cohen, & Norman, 2004a).

Distributed pattern analysis.

As mentioned, the fMRI study described in chapter 2 relies on pattern classification methods detect patterns of brain activity distributed across the brain. In this section, I review the set of studies that first established the viability of these methods for analyzing imaging data. These studies, to date, have focused mostly on the nature of object category representations in the brain.

Studies by Ishai, Ungerleider, and Haxby (2000) and O'Craven and Kanwisher (2000) showed that when a subject is viewing various object categories (face and house), one need not even average over time to detect these events; the co-varying of signal with cognitive state (both perception and

imagery of category) can be seen in the average activity of certain sets of voxels on a second by second basis (O'Craven & Kanwisher, 2000). These areas (such as the fusiform face area and parahippocampal place area) activate strongly and reliably during the perception and imagery of their favored category. However, there seems to also be a constellation of meaningful activation that is either less strong, less reliable, or idiosyncratic to a given subject. Haxby, Gobbini, Furey, Ishai, Schouten, and Pietrini (2001) described a simple correlation-based classifier that is sensitive to the distributed pattern of voxel activities in a brain region. By masking out the maximally active areas of ventral temporal lobe, they showed that there is still useful information in the remaining voxels to make predictions about object category.

While there is still considerable debate about the functional role of this distributed activity in the domain of object representation, several groups of researchers have begun to explore the utility of using pattern classification methods to characterize patterns of brain activity and predict cognitive state (Mitchell, Hutchinson, Niculescu, Pereira, Wang, Just, & Newman, 2004; Cox & Savoy, 2003; Hanson, Matsuka, & Haxby, 2004; Carlson, Schrater, & He, 2003). These researchers have applied a number of classification techniques to fMRI data including correlation (Haxby et al., 2001), linear-discriminant analysis (Cox & Savoy, 2003; Carlson et al., 2003), support vector machines (Cox & Savoy, 2003), a gaussian naive bayesian algorithm (Mitchell et al., 2004), and backpropagation (Hanson et al., 2004; Polyn, Nystrom, Norman, Haxby, Gobbini, & Cohen, 2004b).

1.2.5 Anatomical evidence

The current model attempts to map a number of cognitive systems to a number of anatomical areas. McClelland, McNaughton, and O'Reilly (1995; Norman & O'Reilly, 2003) review the relevant literature regarding the correspondence of the episodic memory system to the structures of the medial temporal lobe. Frank et al. (2001) review the correspondence of a task performance system to regions of PFC and basal ganglia, and I review some evidence for thinking that aspects of contextual targeting reside in PFC as well.

In this section, I review what is known about the pathways between the hippocampal formation

and PFC. For example, many theorists believe that entorhinal cortex acts as the gateway to the hippocampus (Squire & Zola-Morgan, 1991), and surrounding cortical areas such as perirhinal cortex are thought to play a role in familiarity assessment. A number of reciprocal pathways have been discovered linking the hippocampal formation and surrounding cortical areas to areas throughout the prefrontal cortex. While our framework does not yet attempt to ascribe function to these particular anatomical pathways, their general character is suggestive of possible roles. I describe four of these pathways.

First, a medial pathway arises in dorsolateral PFC, courses through the cingulum bundle, and terminates in retrosplenial cortex (BA 30) and the posterior parahippocampal area (Morris, Pandya, & Petrides, 1999; Goldman-Rakic, Selemon, & Schwartz, 1984). Retrosplenial cortex itself has connections with the posterior parahippocampal cortex, the presubiculum and the entorhinal cortex (Morris et al., 1999). The second is a lateral pathway that also arises in dorsolateral PFC, travels in the fronto-occipital fasciculus, and terminates in perirhinal cortex, posterior parahippocampal cortex, and entorhinal cortex (Goldman-Rakic et al., 1984). This pathway provides a direct, single-synapse route from PFC to entorhinal cortex, however, the return projection from entorhinal cortex to PFC is quite sparse.

A third pathway contains inputs from dorsolateral, orbital and medial PFC, and synapses in the medial magnocellular division of the mediodorsal thalamic nucleus (Russchen, Amaral, & Price, 1987). This pathway projects to diverse brain areas, including the entorhinal and perirhinal cortices. The fourth pathway is a direct projection from area CA1 of the hippocampus to prelimbic and medial orbital PFC (Ferino, Thierry, & Glowinski, 1987; Jay & Witter, 1991). This pathway has been shown to support LTP (Ferino et al., 1987).

At present, very few theories of memory are neuroanatomically detailed enough to be seriously constrained by anatomical data. However, when formulating a framework, it is necessary to know that information posited to be shared by two systems indeed has a pathway to get from one system to the other. Once a framework has been set up, anatomical correlates of function can be postulated

and then either verified or falsified.

1.3 Overview of the model

As mentioned above, a major advance in the formalization of models of human memory was developing a framework in which cognitive representations are treated as vectors of scalar values, representing features of a stimulus. Connectionist modelers have taken this approach a step further, by positing that these cognitive representations are activity patterns across sets of neurons in the brain. In the connectionist framework, each element in a representation vector is a unit, and the value of the element is the activity of the unit. Sets of weighted connections determine the interactions of the units in a given network. In section 1.3.1, I describe the forebears to the current model, and the principles underlying the functions of each component in the context of these models. In section 1.3.2, I describe how I have integrated these components, and the rich set of dynamics that arises from this integration.

Given the complexity of the full version of the current model, it is worth setting up at the outset the way the components interact, using the same terms introduced in the previous sections. A schematic of this interaction is shown in Figure 1.5. A contextual representation is maintained in the prefrontal component of the model. This prefrontal representation is perturbed by the presentation of new items in the posterior component of the model. When an item is presented, a gating system is activated, causing subparts of the prefrontal representation to be updated. This causes the prefrontal representation to change slowly over time, much like the context mechanism in the TCM model (section 1.2.2). Item and context representations project into the hippocampal component of the model, creating a conjunctive representation that is then encoded and stored. During the recall process, the hippocampal component is cued with the representation active in the prefrontal component, causing reactivation of stored memories. The details of this story are elaborated in the following sections.

1.3.1 The forebears of the current model

The computational model presented in chapter 4 is an amalgamation of two existing models, originally applied to different domains. Before I embark upon a detailed description of the current model, it is worth taking a moment to review these predecessors as they have been applied in the literature.

The conceptual grandfather of the hippocampal and posterior components of the model is a framework of hippocampal and neocortical interactions elaborated by McClelland et al. (1995). Many of the principles embodied in this connectionist model (such as pattern separation and completion) were developed in this work. These components of the model are more directly based on a model of medial temporal brain structures developed by Norman and O'Reilly (2003), who used it to explain a number of puzzling findings in the recognition memory literature. Specifically, the two interacting components of the model were used to tease apart the influences of a familiarity system and a recollection system on judgments made in recognition memory paradigms. This version of the model contains only the recollection system, as it is likely that the familiarity system has minimal influence during free recall (see section 1.2.1).

Principles of hippocampal function

Critical components of hippocampal function include the ability to create distinct representations for similar stimuli, the ability to form an associative binding of the features comprising an episode and the ability to reinstate learned patterns given a partial cue. The hippocampal model instantiates these functions by incorporating several key features of hippocampal anatomy and physiology as described in O'Reilly and McClelland (1994; McClelland et al., 1995). Figure 1.3 depicts a schematic of the model.

Distinct representations for similar stimuli

As mentioned, patterns from many areas of the brain project to the entorhinal area, which then projects into the hippocampal formation (Squire & Zola-Morgan, 1991). Patterns that are similar

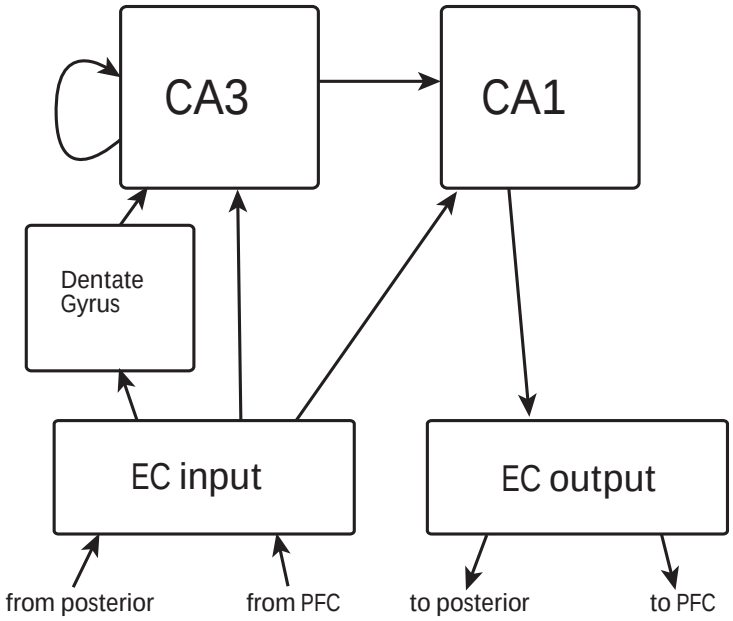


Figure 1.3: A schematic of the hippocampal component of the model.

within the memory system will interfere with each other; this is a major theme of McClelland et al. (1995). The model uses the Dentate Gyrus (DG) layer to perform pattern separation. This sparsely active layer helps to make patterns of activity in CA3 more distinct, even for patterns that are quite similar in entorhinal cortex.

The formation of associative bindings

In this section, I describe the core mechanism responsible for creating the "episode" in our model of episodic memory. As mentioned above, the pattern of activity in hippocampus is determined by both environmental information (the posterior network) as well as internal information (current PFC state). These two sources are bound by means of Hebbian learning mechanisms: Coactive units in entorhinal cortex (EC; the neocortical region that serves as a gateway to hippocampus) are linked to an episodic representation in area CA3 of the hippocampus; connections between active CA3 units are strengthened, and this CA3 engram is linked back to the original pattern of cortical activity via region CA1 (see Fig. 1.3). The strengthening of these CA3 connections forms an attractor, a stable state of the network.

Reinstatement given a partial cue

In dynamical systems terms, each pattern encoded by the memory system becomes a stable fixed point (for our purposes, synonymous with attractor) in the high dimensional space of possible activity configurations of the local network (Strogatz, 1994). When the activity state of the network is similar (in Euclidean space, for example) to one of these stable fixed points, the activity state is drawn towards the fixed point. In cognitive / connectionist terms, these associative bindings have the property that given a partial cue (e.g. the current PFC state), the model can retrieve the entire trace formed at study (the semantic representation and associated PFC state).

There are times, however, when it would be harmful for the model to fall into a particular attractor. For example, during the recall process, if the model fell again and again into the same attractor, very few memories would be recalled. There is a general need, in models of this sort, for a mechanism of transient weakening of previously visited attractors – in more colloquial terms, an

unsticking mechanism.

Many existing models of memory retrieval posit unsticking mechanisms to prevent the model from recalling the same item again and again (inhibition of return; Raaijmakers & Shiffrin, 1981; Becker & Lim, 2003. In the present model, when items are retrieved, a bias weight projecting to the active units becomes slightly more negative. A bias weight represents an input from a tonically active unit; when this weight is negative it acts to suppress the activity of a given unit. Over a short time scale (2-3 trials) the bias weight decays to zero (this is described in more detail in section 4.2.2).

Principles of prefrontal cortex function

What I have referred to as the context maintenance system is based on a model of prefrontal cortex developed by Frank et al. (2001). This work attempts to explain complex task performance in terms of a model that uses active maintenance and gating of representations to hold onto task-relevant information in the face of distracting activity. For this model as well, there is a long and distinguished history, beginning with a model of prefrontal involvement in the Stroop task (Cohen et al., 1990; Cohen & Servan-Schreiber, 1992). The model has been elaborated in a number of studies; the most relevant predecessor to the current version (besides Frank et al., 2001) is described in Braver and Cohen (2000), in which it is shown that a gating model can learn to maintain task-appropriate information using simple reinforcement-based learning rules.

Critical components of prefrontal function include the ability to represent information in a way that is robust to interference and the ability to rapidly update this information as needed (Frank et al., 2001).

The characterization of prefrontal processing has been refined over the years (Miller & Cohen, 2001), but the fundamental story remains the same; prefrontal cortex represents task demands in order to allow humans to flexibly behave in a contextually appropriate manner. The model has continued to evolve in the hands of Frank et al. (2001); it is this more elaborate version of the model that I use.

In the model, PFC is comprised of functional subcompartments called stripes (Pucak, Levitt,

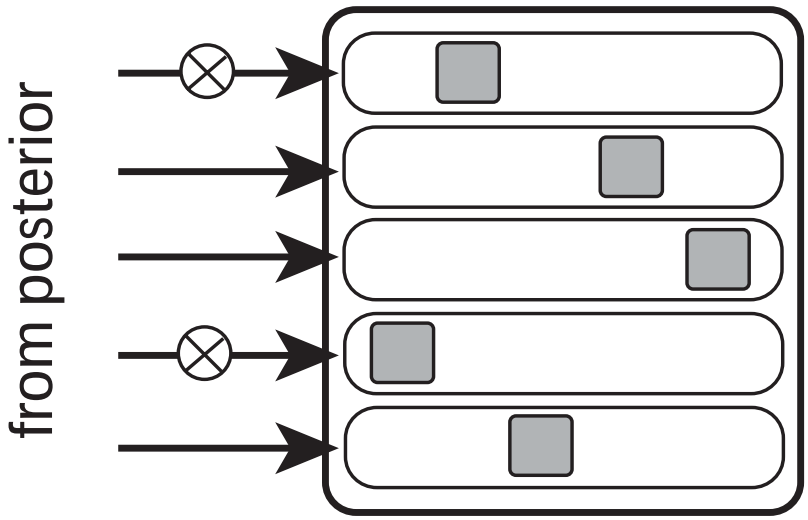


Figure 1.4: A schematic of the prefrontal component of the model.

Lund, & Lewis, 1996), each of which is capable of independently maintaining information over time. Figure 1.4 is a schematic of PFC, showing 5 stripes and their inputs from posterior cortex (described in more detail below). Each stripe is an attractor system with ten stable states; the states can be thought of a high-level representations of item features and task information. In connectionist terms, each stripe is a group of ten units; within-stripe inhibition allows only one (or two) units to be active at a given time (the inhibition algorithm is described in more detail in section 4.2.2).

Frank et al. (2001) also introduced another component to control the gating of the PFC stripes; this controlling system is hypothesized to reside in the basal ganglia system. Specifically, each stripe has a corresponding region in the basal ganglia that determines whether the state of that stripe is "locked" or can be perturbed. The basal ganglia does this by controlling the strength of the self-connection for all units in the stripe. The interaction of the two systems is elaborated below.

Active maintenance

Miller and others have emphasized the ability of prefrontal cortex to maintain activity in the face of distraction (e.g. Miller, Erickson, & Desimone, 1996). Information from the environment (e.g. semantic memory) projects to PFC. Stripes are able to maintain their current state in the face

of distraction through active maintenance: If a unit is active and its self-connection is strong, then this unit will be able to maintain activity in the face of the within-stripe inhibition (see section 4.2.2 for more details). In contrast, activity in posterior areas of cortex (e.g. inferior temporal cortex) is disrupted by distraction (Miller et al., 1996).

Rapid updating

When task demands require, each stripe has the ability to quickly allow new environmental information to be represented and maintained. Each stripe has a gate which controls the strength of the self-connections in the stripe. When the gate is unlocked, the strength of the self-connection is lowered, allowing new inputs to gain access to the stripe. When the gate is locked, the strength of the self-connection is raised, causing the new input to be actively maintained. Functionally, the gate is blocking inputs from disrupting the current PFC state (this is depicted in Fig. 1.4 by an "x" on the posterior projection to the stripe). This locking and unlocking system allows PFC stripes to grab on to features of items in a task-appropriate way. As mentioned, the basal ganglia component of the model implements the gating mechanism by controlling the strength of a given stripe's self-connections. Each sub-region of the basal ganglia can be triggered by afferent inputs. When triggered, the sub-region toggles the state of the corresponding stripe's gate. This mechanism is elaborated in section 4.2.2.

1.3.2 Dynamics of the current model

In this section, I elaborate upon the interaction of the components of the model. The combined model (Figure 1.5) has three components: Posterior cortex, PFC and hippocampus.

Posterior cortex maintains a representation of a studied item, as well as a representation of external environmental features. A word is presented in the external environment, and activates a set of neurons in posterior cortex. The simulations reported in chapter 4 do not implement representations of the external environment in the posterior component.

There are two relevant projections from posterior cortex. First, there is a projection to the PFC.

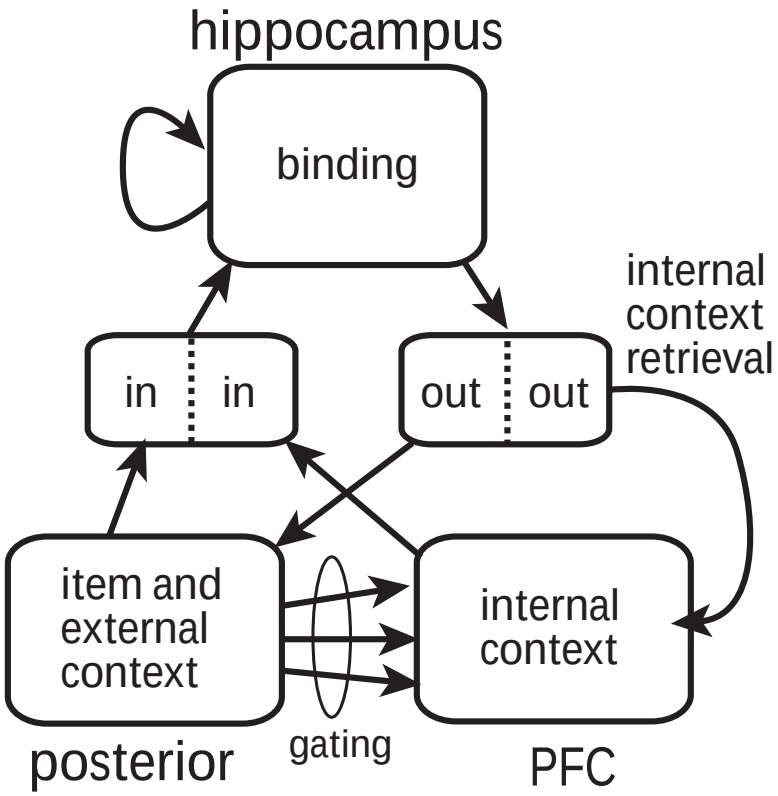


Figure 1.5: A schematic of the tripartite model, showing the set of connections relevant for the current investigations. The posterior system represents item features, PFC represents internal context, and these features are bound in the hippocampus for later retrieval.

During encoding the posterior network projects features of items and the external environment to PFC. Second, there is a projection to the input layers of hippocampus. During encoding, semantic and environmental information are bound up as part of the hippocampal memory trace.

Prefrontal cortex maintains item-related features in the service of task performance. Crucially, PFC projects to the input structures of the hippocampus, such that the PFC state is bound up (along with semantic and environmental information) as part of the hippocampal memory trace for a given episode. At retrieval, this task related information can be reinstated in the PFC (see below) and used to target the appropriate memory traces in hippocampus.

The hippocampus (as described in McClelland et al., 1995; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001) is designed to rapidly encode and store co-active patterns of posterior and PFC activation for later reinstatement. There are two relevant projections from the hippocampus. First, there is a projection from the output layers of the hippocampus to the PFC (Fig. 1.5; internal context retrieval). Retrieved contextual information is loaded back into PFC via this projection, to guide further retrieval attempts. Second, there is a projection from the output layers of the hippocampus to the posterior network. During retrieval this projection reinstates the semantic representations of studied items.

The interaction of the components during memory search

The full model is shown in Figure 1.6. This model can be used to simulate the operations undertaken by PFC to manipulate the memory system. Consider the beginning of the recall period: A number of memory traces lie dormant within hippocampus. The current PFC state is projected into hippocampus as a retrieval cue. This PFC state can be thought of as a spotlight searching for stored hippocampal memory traces (see Figure 1.7). Since this PFC state evolves slowly, it is still shining on the last few items seen, so these items are most likely recalled first (the recency effect; Fig. 1.7(a)). Hippocampus recalls a memory trace, which includes an item and its associated internal context; this internal context represents the state PFC was in when the item was seen during study. The retrieved context is loaded into PFC, re-centering the spotlight on the most recently re-

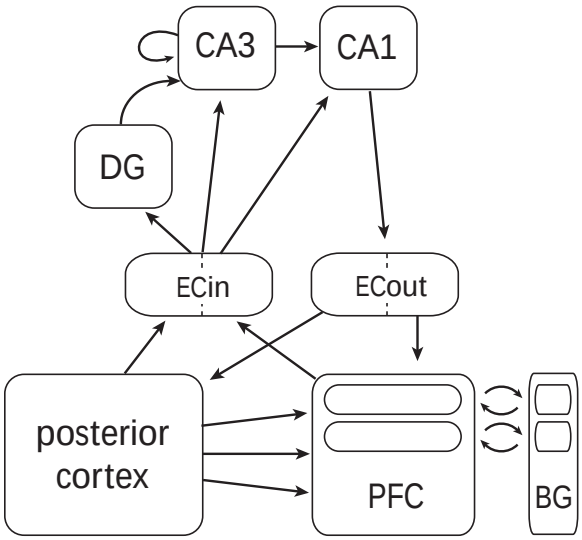
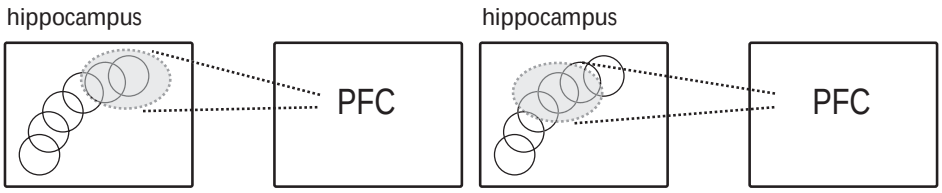


Figure 1.6: A schematic of the full model. PFC = prefrontal cortex, BG = basal ganglia, EC = entorhinal cortex, DG = dentate gyrus, CA1 and CA3 = hippocampal subregions. Not shown are the connections from the posterior layer to the BG layer. See text for details.



(a) At the beginning of the recall period the spotlight shines on the most recently studied items.

(b) When retrieved context updates PFC, the spotlight is recentered on the just retrieved memory trace.

Figure 1.7: The PFC spotlight on memory. Solid circles represent hippocampal memory traces stored over the course of the study episode. The dotted ellipse is the PFC spotlight on memory. See text for details.

trieved item (Fig. 1.7(b); Howard & Kahana, 2002a have a similar theory). In a sense, the retrieved context allows PFC to "jump back in time" (Tulving, 2002) to a state from the study episode. The re-centered PFC spotlight is now a good cue for items that were studied near in time to the previously retrieved item. This cyclic process of using a PFC probe to retrieve memories and using the retrieved information to update the PFC probe allows the model to capture the output patterns made by subjects during free recall.

Given that this is a complex iterative process, I reiterate the major points in outline form. During study, items are presented to the model for encoding. During this encoding period:

- PFC gates in features of these items. Only a subset of the PFC stripes update on any given trial. This causes the updated PFC state to be similar to prior PFC states. This process causes the global state of PFC to drift slowly over time.
- The posterior representation of an item and the PFC representation of internal context both project to the input area of EC, and to hippocampus.
- Hippocampus forms associative bindings of item and internal context information.

During retrieval, there is no external cue to drive recall. Recall is initiated by the internal context still active in PFC:

- PFC is still maintaining the set of features that were present for the final study item (the end of the list state).
- This information is projected to hippocampus as a retrieval cue. Since PFC drifts slowly, the end of list state is a good cue for the last few items in the list.
- In the case that the model successfully retrieves an item and associated context, this retrieved context can be projected back to PFC, making the current PFC state more like one from the past.
- The retrieved posterior representation can drive further update of PFC.
- The newly updated PFC can initiate another hippocampal retrieval. In a sense, the model has jumped back in time (Tulving, 2002). Hippocampus reactivates a prior PFC state and uses that prior state to cue for items.

1.3.3 The model in the context of the literature

In sections 1.2.1 and 1.2.2, I described context as a representation that can be used to target sets of stored memories, and described a series of models that used context vectors to explain behavioral phenomena in human memory. The current model attempts to bridge these more abstract models with the anatomical systems of the brain. While the prefrontal representation in the current model can be thought of as corresponding to the context representation in the TCM model (section 1.2.2), there are important structural differences between these models. The nature and implications of these differences will be more fully explored in chapter 4, in the context of a series of simulations of young and older subjects' performance in free recall (also see section 1.2.2).

The current model emphasizes that the prefrontal component can be used as context to target stored memories. However, in section 1.2.1, I argued that many types of representations could be used to target memories, including representations of the external environment and stimulus category, which may not reside in prefrontal cortex. It is reasonable to think that representations

from many areas of the brain can be used to target stored memories. However, the model does posit that prefrontal cortex has specialized machinery that allows it to maintain a context representation in the face of irrelevant activity (section 1.3), such as that presented by a distractor task in a free recall paradigm. Thus, the prefrontal cortex may have a special involvement in the memory search of the free recall paradigm, which is consistent with the finding that damage to prefrontal cortex disrupts memory search (section 1.2.3).

In section 1.2.1 I explored the concept of inner mental context as it relates to the directed forgetting paradigm. Given the description of the current model in section 1.3, I can now define inner mental context more rigorously. In the current model, the PFC representation is modified as item representations are processed by the system. Thus, the PFC representation acts as inner mental context by maintaining an amalgamation of recent states of the system. This framework permits a mechanistic explanation of both the effect of distractor activity during free recall, and the effect of a forget instruction in directed forgetting. The simulations presented in chapter 4 only begin to explore the phenomena outlined in this chapter. In chapter 5, I outline how the model may be used to fit behavioral performance in a number of variants of the free recall paradigm.

1.4 Conclusion

In this chapter I have introduced the concepts common to the three investigations described in the rest of the dissertation. In chapter 2, I describe an fMRI study of free recall. I argue that patterns of brain activity detected in this experiment are being used to target stored memories. In chapter 3, I describe a behavioral study of free recall. In this study I manipulate the encoding task during the study period in an attempt to perturb the subject's task context representation. Items studied in the context of different tasks are not recalled nearby in time (relative to a control condition), suggesting that these items are isolated from each other in memory. In chapter 4, I describe a set of simulations of the free recall paradigm using the model presented in section 1.3. By perturbing the prefrontal representation during the study period, I can qualitatively fit the basic results of the behavioral study

in chapter 3. Finally, I describe a series of simulations in which I damage various components of the model, in an attempt to fit the free recall performance of the older subjects described in section 1.2.3. Finally, in chapter 5, I elaborate upon the future directions for each of these investigations, and the ways in which they inform each other.

Chapter 2

A neuroimaging study of free recall

2.1 Introduction

Tulving likens the process of episodic memory to time travel (Tulving, 2002). In recalling the past, we revisit details and features of events no longer present in the environment. This type of memory is termed episodic because it is a binding of all types of stimuli, both external and internal, into a single episode, or event. In this chapter, I investigate the process by which the brain revisits past episodes.

In the first chapter I reviewed theories of context in memory search, both from a behavioral (section 1.2.1) and modeling (section 1.2.2) perspective. In this chapter, I interpret the results of an imaging study of free recall in terms of those theories. According to these theories, context representations present in the cognitive system when memory traces are encoded can be used later to cue the memory system to revive these stored traces. The presence of a context representation in the system will bias the memory system to retrieve traces congruent with that representation. This context representation is being used to target a set of memories (see section 1.2.4 for a full exposition). Thus, in a free recall experiment, a pattern of brain activity involved in memory targeting should be present during the study period, reinstated during the recall period and its strength should correlate with the recall performance of the subject. In section 2.3 I describe patterns of brain activity that

meet all of these criteria.

In the free recall paradigm described here, the study items were drawn from three distinct categories (famous faces, famous locations, and common objects), providing the subjects with a set of easily accessible memory cues. These categorical items also act as a sort of 'contrast dye', as previous studies have shown that each category elicits a distinct neural signature (Haxby et al., 2001). As shown in section 2.3.3 patterns of brain activity fall into a characteristic state during each of the three study tasks. Certain voxels tend to increase their signal, while others decrease. A backpropagation classifier (section 2.2.10) is used to characterize each of these three brain-wide patterns. Furthermore, this classifier provides a means of quantifying, during the recall period of the experiment, the degree of match between the current brain pattern and each of the characteristic study patterns (sections 2.3.4 and 2.3.5). The re-emergence of these characteristic patterns during the memory search of free recall turns out to be predictive of the subject's behavior. Specifically, one can predict with high accuracy the category of the word that is recalled by the subject. A final set of investigations compares the results generated with the classifier to results generated using a simpler approach, in order to assess the benefit of using this type of multivariate analysis (section 2.3.9).

2.2 Methods

2.2.1 Overview

Subjects (section 2.2.2) were run in a free recall paradigm (sections 2.2.3 and 2.2.4). Verbal responses made by the subject during the recall period were recorded and scored (section 2.2.5). During the entire experiment, signal was acquired from their brains using an MRI scanner (section 2.2.6). The data was preprocessed, and whole-brain masks were created (section 2.2.7). Two more sets of masks were created, based on the subjects' anatomy (section 2.2.8), and based on statistical tests of the study period data (section 2.2.9). Pattern classification methods were applied to each of these masked datasets, both to characterize the consistency of the patterns of brain activity seen during the study period, and to determine the degree of correspondence between the classifier

output and the verbal responses made by the subject (sections 2.2.10 and 2.2.11). An event-related average was constructed to investigate the onset of classifier activity relative to the verbal responses made by the subject (section 2.2.12). Finally, a set of non-parametric statistics were applied to the data to determine the significance of the results (section 2.2.13).

2.2.2 Subjects

Fourteen subjects were run in the experiment (9 female; ages 19-27). Two subjects were excluded from the final analyses because of failure to follow instructions, one was excluded because of technical difficulties with the recording apparatus, one was excluded because of excessive movement in the scanner, and one was excluded because they asked to be removed from the scanner. This left the nine subjects presented here. Informed consent was obtained in a manner approved by Princeton IRB.

2.2.3 Materials

Materials consisted of color and black and white photographs drawn from three categories: celebrity faces, famous locations or landmarks, and common objects. Photographs were collected from free sources on the internet, and were chosen to be distinctive and memorable. Over the entire experiment, subjects saw thirty stimuli from each category.

Examples of the celebrity faces are Bruce Lee, Brad Pitt, Halle Berry and Madonna. Examples of the famous locations are the Taj Mahal, the Saint Louis Arch, Epcot Center and Graceland. Examples of the common objects are a colander, tweezers, a zip-lock bag and spray paint.

During the study period of the experiment, these photographs were presented with a name written in text above them (for example, a picture of Jack Nicholson, with the words "Jack Nicholson" above the picture). Subjects were asked to remember this name for the later recall periods. The tasks performed on these items during the study period are described below.

The experiment was presented using E-Prime 1.1 (Psychological Software Tools, Pittsburgh, PA) run on a Windows PC. The PC was connected to a projector, which projected the images onto

a screen behind the supine subject. This screen was viewed by the subject through a periscope-style mirror placed near their eyes.

2.2.4 Behavioral procedure

The general design of the experiment follows Watkins and Peynircioğlu (1983), with a few key differences. Each experimental block consisted of one study period followed by three recall periods. After three experimental blocks of this sort, there was a final block which consisted of a single final free recall period. All of the experimental blocks were preceded by a practice period in which subjects practiced both the stimulus judgments and the arithmetic task (see below). Each period of the experiment is described in detail below. All periods of the experiment took place in the scanner, while the scanner was acquiring functional images.

The study period

Each study list was composed of thirty study trials. Each study trial consisted of a cue period (1.8 sec), a stimulus period (4.5 sec), a judgment period (2.7 sec) and a gap (1.8 sec). During the cue period, a title screen appeared, orienting the subject to the category of the upcoming stimulus. The three titles were "Judge the famous", "Vacation time", and "Object lessons". During the stimulus period, the photograph and name of the stimulus were presented on the screen (section 2.2.3), and subjects were asked to prepare a response for the upcoming judgment. During the judgment period, subjects made a rating that was specific to the category of stimulus presented. For the faces, subjects made a rating of love or hate for the celebrity; for the locations, subjects rated how much they would like to visit the location; for the objects, subjects rated how often they come across it in their day-to-day life. All judgments were made by a key-press on a stimulus response glove (Five-button IFIS response glove, Psychological Software Tools, Pittsburgh, PA) on a scale from 1 to 5.

After each study trial, subjects performed an arithmetic task (12.6 sec). During the arithmetic task, subjects saw two equations (example: 14 + 15 + 5 = 35) and were asked to make a true/false response. These responses were also made with a key-press on the glove. Subjects were given a

few seconds to respond to each equation. If they waited too long to respond, they were given a 'LATE' message, and the next equation was presented. Subjects were instructed that the arithmetic equations were meant to be challenging, and were asked to attempt to solve each problem.

During each study period, subjects were exposed to 30 items, consisting of 10 items from each category, presented in a pseudo-random order; every set of three items contained an item from each category, but order was randomized within the sets of three.

Recall by category

At the end of each study list were three recall periods, each of which lasted 54 seconds (this was during the same scanner run). During each of the three periods, subjects were instructed to report as many items as they could remember from the current list, from a particular category. For example, a subject might first get instructions to recall words from the "Judge the famous" task. Subjects were asked to recall words from each category in turn; as soon as one period ended, there were instructions orienting the subject to the next recall category. Within-category, subjects were told to recall the items in any order they pleased. Subjects were asked to continue to attempt to recall for the entire recall period. The order of category-recall was randomized across runs.

Final free recall

After the third run, there was a final recall period, lasting 162 seconds. Subjects were told at the beginning of the experiment that there would be four runs of the memory experiment. At the beginning of the fourth run subjects informed of the nature of this final test. They were told that this final run would only last about three minutes, and that they would be asked to recall as many items as they could from the entire experiment, in any order they pleased. Subjects were asked to continue to attempt to recall items for the entire period.

2.2.5 Behavioral analysis

Verbal responses were acquired from the subject in the noisy environment of the scanner bore. A funnel was placed near the mouth of the subject, attached to a length of plastic tubing. At

the end of the tubing was a small microphone which was connected to a Macintosh PowerBook G4 laptop in the scanner control room. Each run was recorded in its entirety using Audacity (`http://audacity.sourceforge.net`, ©2005 Audacity development team, distributed under the GNU General Public License) an open-source audio editor and recorder available freely on the internet. To reduce the background noise of the functional scans being acquired during the recall period, a digital noise reduction algorithm was applied to the signal. The algorithm ('Noise Removal', bundled with Audacity) was initialized with a 5-second block of the file, containing only scanner noise. Then the algorithm attempted to remove the scanner noise from the entire signal, improving comprehensibility significantly. The recall periods were then analyzed using specialized software designed for parsing free recall response data (Parse, maintained by the Kahana lab; `http://memory.psych.upenn.edu`). The Parse software recorded the millisecond onset and item reported for each verbal response. Given the millisecond onset times of each response, I divided the run into 1.8 second intervals (the time for a full brain scan to be acquired), and assigned each response to a scan bin. If a response was initiated during a particular 1.8 second bin, then it was assigned to that bin. If a subject was recalling quickly, it was possible that two responses could be assigned to the same bin. As such, two responses from different categories could be assigned to the same bin.

I generated three vectors for each recall period, one corresponding to responses from each category. An integer was assigned to each element of the vector, corresponding to the number of recalls made from that category during that particular brain scan. I refer to this set of vectors as the "recall record". These vectors were used to assess the degree to which the classifier output corresponded to the recall behavior of the subject (section 2.2.10).

Behavioral measures of subject performance were also generated from these vectors. Table 2.1 reports the percent recall for each category, and average recall across all categories, for both the recall-by-category periods, and the final recall period. The standard error reported is calculated across subject means.

2.2.6 *Imaging methods*

This experiment was run on a Siemens Allegra 3 Tesla scanner housed at the Psychology Department at Princeton University, and maintained by the Center for the Study of Brain, Mind and Behavior. Each subject received seven scans, one scout to ensure proper placement, an MP-RAGE structural scan lasting about 10 minutes, a test functional EPI scan to ensure proper brain coverage, and EPIs for the four experimental runs described above in section 2.2.4.

The MP-RAGE structural scan had whole-brain coverage. 176 sagitally oriented slices were acquired. Relevant parameters: TR = 2500msec, TE = 4.38msec, voxel size = 1.0 by 1.0 by 1.0mm, flip angle = 8^o, FOV = 256mm.

During the functional EPI scans 30 oblique slices covered the whole brain. The slices were oriented through the AC-PC line. Relevant parameters: TR = 1800msec, TE = 30msec, voxel size = 3.0 by 3.0 by 3.6mm, flip angle = 75^o, FOV = 192mm.

The first three experimental runs consisted of 510 scans, while the final run consisted of 104 scans.

2.2.7 *Preprocessing the imaging data*

Preprocessing of the functional data was done with the AFNI software package (Cox, 1996; Cox & Hyde, 1997; available at `http://afni.nimh.nih.gov/afni`; this package includes the algorithms 3dvolreg, 3dDespike and 3dDetrend, mentioned below). The first five scans of each run were discarded to allow scanner signal to stabilize. After this discard, all functional volumes were registered to the first functional scan from run 1 using the motion correction algorithm 3dvolreg (Cox & Jesmanowicz, 1999). Signal spikes were then removed from the data using 3dDespike, followed by removal of mean, linear and quadratic trends using 3dDetrend. Finally, the functional data was spatially smoothed using a Gaussian blur (full-width half-max of 4mm).

The functional data were then loaded into Matlab (Mathworks, Natick MA), using a combination of AFNI-bundled Matlab scripts and a Matlab-based software package designed by the author. The

time-course of each voxel was replaced by a z-score normalized version. Z-scoring was applied on a voxel-by-voxel basis (that is, across time) within run. This z-score normalization was meant to remove differences in voxel baseline across runs, and normalize variance across voxels. Finally, I applied a feature selection process to the data, which reduced the number of voxels analyzed by the classifier by nearly an order of magnitude. The feature selection process was used to select voxels whose signal significantly deviated between study conditions. An ANOVA was run on a voxel-by-voxel basis, looking only at study period data (the three groups were defined by the three study tasks). If a voxel showed a p-value significance < 0.05, it was included in the classification analyses. As such, a different number of voxels was analyzed by the classifier for each subject. These voxel counts, by subject were: 6482, 10008, 9897, 7809, 8311, 4517, 3951, 6661, and 7826. For example, each training and testing pattern for subject 1 was composed of 6482 voxels.

2.2.8 Creation of masks of anatomical subregions

Classification analyses were run on whole-brain data, as well as on datasets restricted to particular anatomical subregions. These focused analyses were carried out because classification analyses on the whole-brain data suggested that temporal lobe structures were more strongly activated during the encoding period. As tables 2.10 and 2.12 show, there is recall-related activity in both frontal and temporal areas, but the temporal lobe pattern seems to be stronger and more reliable.

These anatomical masks were created using the AFNI software package, specifically the set of Talairach plugins, including a detailed map of brain regions expressed in Talairach coordinates (see section 2.2.7 for details of the AFNI package). A Talairach transformation was applied to each brain. Two anatomical masks were created (frontal lobe and temporal lobe) in Talairach space. I decided to apply a liberal criterion as to which brain regions to include in each mask. The frontal mask consisted of the following regions (as named in AFNI): precentral gyrus, superior frontal gyrus, medial frontal gyrus, middle frontal gyrus, inferior frontal gyrus, orbital gyrus, rectal gyrus and anterior cingulate. The temporal mask consisted of the following regions: transverse temporal gyrus, inferior temporal gyrus, middle temporal gyrus, superior temporal gyrus, parahippocampal

gyrus, fusiform gyrus, uncus, and hippocampus. These masks were then transformed into the original anatomical space of each subject (before the Talairach transformation) and loaded into Matlab. The same feature selection process was applied to the voxels in each of these masks, as described in section 2.2.7. Thus, the set of voxels in each of the anatomical masks is a subset of the voxels in the whole-brain mask.

2.2.9 Creation of ANOVA contrast masks

The core results of this chapter spring from classification analyses carried out by training a backpropagation-based classifier on data from the study period of the experiment. I also ran similar analyses using a more traditional approach (described below) to validate the use of these classification methods.

I generated a set of three masks, one for each of the stimulus categories. Each mask was constructed using an ANOVA contrast on the study period data. The three contrasts looked for voxels that were significantly activated by one of the three study conditions and were constructed as follows: [2,-1,-1],[-1,2,-1],[-1,-1,2].

There were equal observations in each of the three study conditions, thus, effect size and F-value were equivalent measures of degree of response of a given voxel. Each mask consisted of the set of 50 voxels with the largest F-values for that contrast. No contiguity requirements were considered for inclusion in the mask. An average time series was constructed for the set of voxels in each mask, as follows: signal from the 50 voxels in each mask was averaged together, to obtain three time series for the final recall period. These time series were compared to the verbal recalls made by the subjects as described in sections 2.2.11 and 2.2.12.

Table 2.14 and Figure 2.5 show the results of the analyses on these averaged time series. They do not correspond to the behavior of the subject nearly as well as the backpropagation classifier.

2.2.10 The backpropagation-based classifier

A network classifier was used to demonstrate the correspondence between patterns of activity present in the data and the verbal recalls made by the subject. This network classifier was implemented with the Matlab Neural Network toolbox (Mathworks, Natick MA). A schematic of a backpropagation network is depicted in Figure 2.1.

The classifier was trained using the conjugate gradient descent variant of the backpropagation algorithm ('traincgb' in Matlab). A two-layer network was used. The input layer contained one unit for every voxel that passed the feature selection process (section 2.2.7). The output layer contained three units, one for each stimulus category (faces, locations, and objects). The connections in this network were purely feedforward (with full connections from the input layer to the output layer). The output units used a sigmoid transfer function ('logsig' in Matlab). A cross-entropy function was used to calculate network performance during training. Equation 2.1 shows this cross-entropy function: t is a vector containing the target values for each output unit, y is a vector containing the output of the classifier and p is the measure of performance.

$$p = -\sum \left(\sum \left(t * \log y + (1 - t) * \log \left(1 - y \right) \right) \right) \tag{2.1}$$

Bishop (1995) describes the advantages of using a set of backpropagation networks to reduce prediction error. In all analyses involving the backpropagation classifier, a given output trace is an average output of 50 networks.

The backpropagation algorithm is used to set the weights of the network such that the network minimizes the difference between the target values and the network outputs. There are a number of excellent references available containing a thorough description of backpropagation network classifiers, at an algorithmic level (Bishop, 1995; Duda, Hart, & Stork, 2001; Rumelhart, 1996) and at a practical level (LeCun, Bottou, Orr, & Müller, 1998).

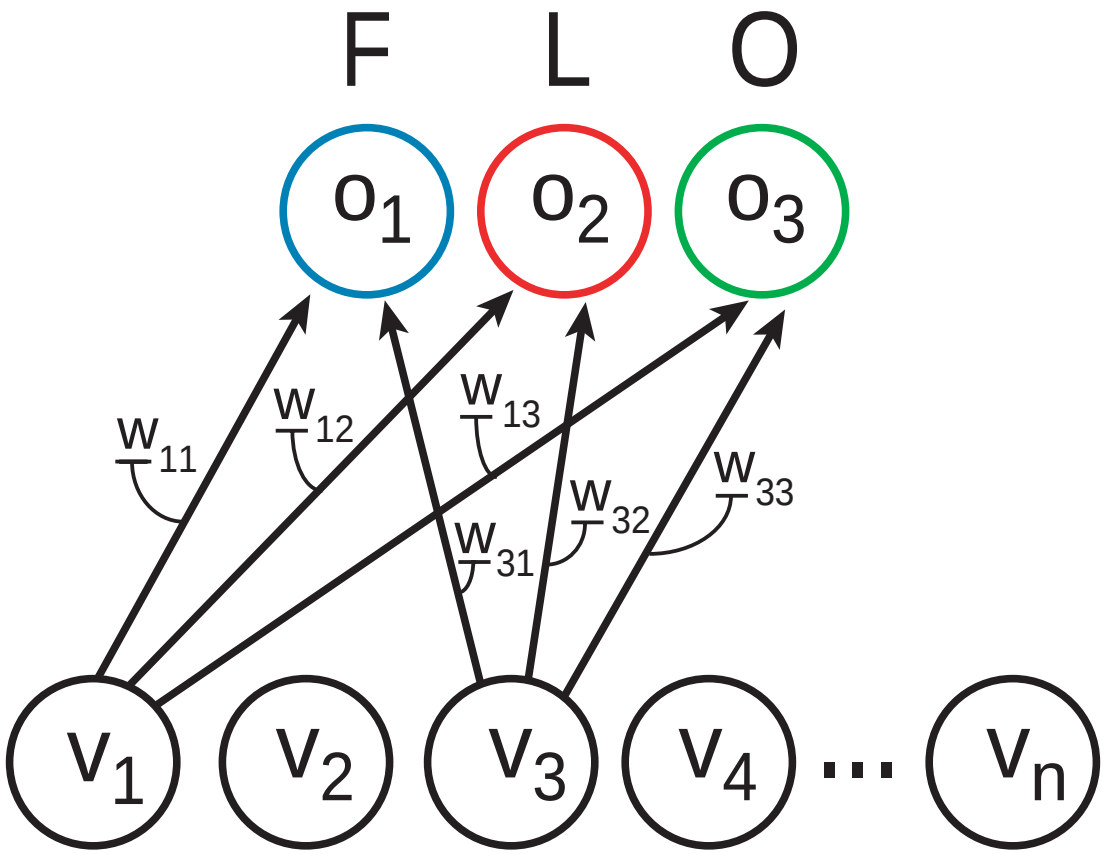


Figure 2.1: A schematic of a backpropagation-based network classifier. v_i = input unit with signal from voxel i; o_i = output unit corresponding to category i; w_{ij} = weight connecting input unit i with output unit j. The backpropagation algorithm is an error-driven learning algorithm. During training, the network calculates an error signal based on the difference between the actual activity pattern at the output layer, and the target activity pattern for the output layer. This error signal is used to determine the direction of weight change in the network. Weights are changed such that the next time a pattern is presented, it will evoke a smaller error signal.

Details of training

The network is trained on the preprocessed imaging data described in section 2.2.7. On each study trial subjects are presented with a stimulus drawn from one of three categories. Studies of the brain response to environmental stimuli suggest that observable changes in the blood flow of the brain do not appear in fMRI signal until a few seconds after stimulus onset, with the peak of this signal occurring 4 to 6 seconds after stimulus onset (Huettel, Song, & McCarthy, 2004). As such, I train the network on the set of 5 consecutive scans starting on the third scan after stimulus onset, for each study trial. This relatively crude accounting of the brain's hemodynamic fluctuations yields us 450 training patterns across the three study periods of the experiment (section 2.2.4). Each training pattern has a corresponding target pattern for the three-unit output layer of the network. This target pattern is binary; the target value corresponding to the studied category is set to one, and the other two target values are set to zero.

Before training, the connection weights of the network were initialized to random values. The presentation order of patterns during training is randomized. Training was stopped when either the mean cross-entropy error across the output layer fell below 0.001 or the network made 500 passes through all of the training patterns. Preliminary investigations suggested that the exact values of these parameters were not critical to network performance.

Testing the classifier

After training, the weights of the network are fixed, and the network is presented a set of brain patterns drawn from some portion of the experiment that it was not trained on. For each test pattern, the output layer of the network produces three activation values. These activation values can be interpreted as a measure of how much the current test pattern matches the characteristic brain state of each of the study tasks. The next sections deal in more detail with issues of interpretation of the structure of this trained network and the output values it produces.

Generalization tests on the study period data

A leave-one-out or generalization analysis (Duda et al., 2001) was conducted on the training patterns drawn from the study periods of the experiment (see above). A backpropagation network, of the type described above, was trained using study period data from two of the runs. I then tested this network on the study period data from the third run. By iterating through all three runs (each time leaving aside a different run's study period for testing), one gains a sense of how consistent the category-related brain activity patterns are across runs. Each iteration of this algorithm starts with a freshly initialized network. The results of this analysis are presented in Table 2.2. This analysis serves to validate that the classifier is detecting consistent patterns of activity during the study periods.

Scoring the classifier during recall-by-category

The current experiment has two recall phases (section 2.2.4), three in which the subject performs recall-by-category, and one in which the subject recalls across categories. I developed a simple measure of classifier performance during the recall-by-category periods. I took the set of time-points on which a subject made a correct recall, then calculated the average activity of each of the classifier output units over this set of time-points. Table 2.3 describes these results. The recall responses were forward shifted by three scans to account for lag to peak hemodynamic response (see *Details of training*, above).

Application of the classifier to final recall data

Figure 2.2 shows the output of the classifier over the course of the final free recall period for a single subject. During the first eight scans, subjects read instructions describing the final free recall block (section 2.2.4). The recall responses were forward shifted by three scans on this graph, to account for lag in the peak hemodynamic response to an event (see *Details of training*, above). The size of the circle represents the number of items recalled on a given scan (smaller circles represent a single recall, larger circles represent two or more recalls). The three lines were temporally smoothed for illustrative purposes. However, all analyses were performed on unsmoothed data. To perform the

smoothing, each point on the graph was replaced by an average of it with the immediately preceding and succeeding time-points. Section 2.2.11 below describes the methods by which I quantified the correspondence between the classifier output traces and the verbal recall record for each category.

Interpreting the weight structure of the network.

Figure 2.4 shows three classifier-derived colormaps, one for each of the three study categories. These colormaps depict the set of voxels that passed the feature selection process; the color code reflects the relative importance of a given voxel on the output unit for that category.

A simple algorithm is used to determine the level of influence of a given voxel to a given category. The value produced by the algorithm is the 'importance value' of the voxel. In the context of the network classifier, an input unit contains signal from a single voxel. The activation value of an input unit can be positive or negative. The state of an output unit is determined by a weighted sum of all of the input units, passed through a sigmoid function (see Fig. 2.1). As such, the activation value of an output unit can only range between zero and one.

$$imp_{ij} = w_{ij} * a_{ij} \tag{2.2}$$

Equation 2.2 describes the calculation of this importance value. The subscript i is used to step across input units in the network. The subscript j is used to step across output units in the network. Thus, a single input unit has an importance value for each category. This importance value is defined as the product of two values, the weight between input unit i and output unit j, and the average activity of input unit i for the training patterns drawn from category j.

The importance value is meant to reflect how effectively this input unit can alter the value of the corresponding output unit. The sign of the importance value reflects whether activity in this input unit tends to drive the activation value of the corresponding output unit up or down. If both w_{ij} and a_{ij} are positive (or both are negative), signal in the input unit causes the output unit to turn on; it has a positive importance. If the two have opposite sign, then the input unit's signal acts to turn off

this output unit, and it has a negative importance value.

Network classifiers use both positive and negative evidence to reduce error on the training set (Polyn et al., 2004b). For example, if the signal in a voxel increases when a face is viewed, then a strong positive weight to the face output unit will reduce error. A strong negative weight from this input unit to the location output unit will also reduce error, by turning off the location output when face is the correct category.

Generating brain maps

The importance values described above were used to create the brain maps shown in Figure 2.4. The structural images of the nine subject's brains were transformed to Talairach space using the AFNI software package (this same transformation is used in section 2.2.8). This provided the subject-specific transformation necessary to convert each subject's colormap into a common space. For each subject, each voxel was assigned an importance value for each category, described in detail above. The color of each voxel represents the mean value of importance over all nine subjects.

The colorscale represents the importance of the voxel relative to the other voxels in the network (see above). Positive values are shown in red, yellow and green, in decreasing order. Negative values are shown in dark green and blue, in decreasing order. If the absolute value of a voxel's importance value was less than 0.0041, that voxel was not included in the map. The structural image from subject 1 was used as an underlay for the montage, which was created with the AFNI software package (Cox, 1996).

2.2.11 Correlations and percent correct

In this study I compare the results of several classification analyses to each other. In order to do this, I describe two metrics for quantifying the correspondence between the output of a given classifier and the record of verbal responses made by the subject during the final recall period. The first metric was a correlation. This measures the degree to which a given output unit activation trace covaries with the recall record for a given category (section 2.2.5). All pairwise correlations between

the three output traces and the three recall records were calculated (see, for example Tables 2.4 and 2.6). The second metric calculated percent correct (see Table 2.8). For this metric, I excluded single scans on which the subject made verbal recalls from more than one category. The verbal recall record was shifted forward by three time-points to account for lag to peak hemodynamic response (see section 2.2.10). A response was deemed correctly identified if the activation value of the output unit corresponding to the category of the item recalled was the maximum of the three activation values. Significance of the correlation values and percent correct scores was computed using a non-parametric method, described in section 2.2.13.

Section 2.2.9 describes a method for generating category-related average activity traces. These activity traces were subjected to the same correlation and percent correct analyses as the backpropagation traces, as shown in Table 2.13.

The classifier, trained only on study data, is sensitive to the category of the item being recalled by the subject. Each correlation, described above, only characterizes the sensitivity of the classifier to a given category. In order to quantify the overall correspondence between classifier and behavior across all three categories, it is necessary to collapse this three-by-three correlation matrix to a single number. Here, I describe a simple metric that quantifies this correspondence. I calculated the average difference between the on-diagonal and off-diagonal elements of the correlation matrix; this number is referred to as the OnOff metric. Table 2.7 describes the results of applying this metric to each subject's data.

2.2.12 The event-related average

The output traces of the classifier allow one to visualize the second-by-second fluctuations in strength of the study-related brain patterns. In order to characterize the average fluctuation of pattern strength relative to a verbal recall event (regardless of the category recalled), I created an event-related average of classifier output.

As described in section 2.1, I hypothesize that during the period preceding a successful verbal recall a cue-construction process is underway. This predicts that the activation value of the output

unit corresponding to the recalled category should rise before the recall is verbalized. As such, I exclude the set of recall events where there has been a same-category recall in the previous 8 scans (14.4 seconds). Thus, the set of events included in the event-related average are verbal recalls from a given category where there have been no same-category recalls in the prior 8 scans. This large window is necessary to ensure that hemodynamic brain activity related to a prior recall event will have had a chance to return to baseline.

For each recall event that is used in the event-related average, the output trace corresponding to each stimulus category is assigned to one of three pools: Currently Recalled, the category of the item reported at time zero (black line); Recently Recalled, any category that has been reported in the last 8 scans (red line); Not Recently Recalled, any category that has not been reported in the last 8 scans (purple line). By design, no category can fall into more than one bin for a given recall event.

For example, if one considers the recall event where the subject reports 'Jack Nicholson', and has just reported 'Taj Mahal', but nothing else in the last 8 TRs, then the output values from the face unit will be averaged with the Currently Recalled traces, the output values from the location unit will be averaged with the Recently Recalled traces, and the output values from the object unit will be averaged with the Not Recently Recalled traces.

The y-axis of the plot represents average classifier activity over the set of events. Each point on the graph is an average of nine subject means. The sub-graph beneath the event-related average shows the difference between the Currently Recalled and the Not Recently Recalled traces at each time-point. A paired t-test is calculated on the set of mean differences across the nine subjects at each time point. The error bars represent the standard error on the set of difference scores.

An event-related average is constructed for the output of the whole-brain classifier (Figure 2.3) and for the average traces generated by the ANOVA contrasts (Figure 2.5, see section 2.2.9 for details).

2.2.13 Non-parametric statistics

Non-parametric statistics seem most appropriate to judge the significance of results obtained in the current experiment. These non-parametric methods were used at four points in the experiment, to determine the significance of the individual elements of the correlation matrix for each subject, the significance of the OnOff correlation measures for each subject, the significance of the set of nine OnOff measures across the entire experiment, and the significance of the percent correct numbers calculated for each subject.

I used a wavelet-based signal decomposition procedure (using Matlab; Mathworks, Natick MA) to generate surrogate classifier output time-courses that had the same spectral characteristics as the original time-courses. This procedure was based on methods described by Bullmore, Long, Suckling, Fadili, Calvert, Zelaya, Carpenter, and Brammer (2001). I used a Daubechies wavelet to decompose the classifier output time-courses, and then scrambled the coefficients within-scale. By reconstructing a waveform with the scrambled coefficients, I obtained a surrogate time-course with similarly scaled fluctuations as the original (Bullmore et al., 2001).

Each output unit received its own surrogate data set. That is, I decomposed the time-course of each output trace separately, to create a set of surrogate time-courses that matched that particular output unit. For each output unit I created a set of ten thousand surrogate time-courses (giving a full set of thirty thousand). By picking one surrogate time-course from each set, I was able to reproduce the two major analyses of the experiment, the three-by-three matrix of correlations and a percent correct measure.

I created a distribution of OnOff correlation metrics and a distribution of percent correct measures, derived from the surrogate data. By comparing the actually obtained data to these distributions, I was able to determine the p-value: the probability that, by chance, a surrogate metric exceeded that obtained in the experiment.

An alternate non-parametric statistic was devised for determining the significance of the across-subject results. For this statistic no surrogate time-courses were generated. Each subject had a

three-by-three correlation matrix used to calculate the OnOff metric. By scrambling the labels of the columns of the correlation matrices, and recalculating the set of nine subjects' OnOff metrics, I tested the probability of obtaining this set of OnOff measures by chance. Ten thousand relabelings were performed; for each, the columns of each of the nine subjects' correlation matrices were relabeled, a new set of nine OnOff metrics were calculated, and the mean of these nine metrics was calculated. This set of ten thousand mean OnOff metrics was used to generate a p-value for the actual result obtained in the experiment, and is reported in Table 2.7 (bottom panel).

2.3 Results

2.3.1 Overview

First, the behavioral performance of the subjects in the various recall conditions of the experiment is presented (section 2.3.2). Then the results of the generalization analysis performed by the backpropagation classifier (section 2.2.10) are presented, in order to assess the stability of the patterns detected by the classifier during the study period, across the first three runs (section 2.3.3). Then, the results of the recall-by-category analysis are presented. This section shows that the classifier output unit corresponding to the category being recalled tends to be the most active during this recall period. A series of analyses on the final recall data are then presented (section 2.3.5), applying the backpropagation classifier to this data. In this section, correlation and percent correct metrics are presented. In section 2.3.6, an event-related average of classifier output is presented, to investigate the onset of the brain signal relative to the recall verbalization. Then, in section 2.3.7, a map of the importance values generated from each subject's trained backpropagation classifier is plotted over an anatomical scan of the brain. In section 2.3.8, I present the results of applying the backpropagation classifier to frontal and temporal lobe anatomical subregions. Finally, in section 2.3.9, the average time-courses of the ANOVA contrast masks (described in section 2.2.9) are used to create correlation metrics and an event-related average, to assess the advantage of using the backpropagation classifier.

	Recall category			
	Face	Location	Object	Across
Runs 1-3	0.59 (0.03)	0.52 (0.03)	0.40 (0.03)	0.51 (0.04)
Run 4 (Final)	0.50 (0.03)	0.43 (0.04)	0.27 (0.03)	0.40 (0.02)

Table 2.1: Mean recall performance over all subjects, broken down by category (columns one through three) and averaged over category (column four). Standard error over subject means is in parenthesis.

2.3.2 Behavioral results

Table 2.1 presents percent recall by category and across category, averaged over all subjects. Recall performance is presented separately for the recall-by-category periods (Runs 1-3) and the final recall period (Run 4). Subjects, on average, recall more items in the initial recall periods than in the final recall period. Furthermore, the different stimulus categories showed different levels of recall, with face being the greatest, followed by location and then object.

2.3.3 Generalization analysis on the study period data

Section 2.2.10 describes the generalization analysis presented in Table 2.2. For this analysis a backpropagation classifier was trained on the study period data from two of the runs, and was tested on the study period data from the third run. This procedure was repeated three times, once for each left-out run. The percent correct performance was then averaged over the three runs. For this analysis, chance performance is 0.33.

2.3.4 Analysis of recall-by-category periods

Section 2.2.10 describes the analysis procedure for the recall-by-category periods in runs 1-3. These results are presented in Table 2.3. Note that the on-diagonal elements are larger than the off-diagonal elements, indicating that the classifier is sensitive to the identity of the category being recalled.

Subject	Face	Location	Object	Average
1	0.91	0.76	0.78	0.82
2	0.91	0.85	0.73	0.83
3	0.90	0.62	0.55	0.69
4	0.87	0.74	0.71	0.78
5	0.89	0.54	0.51	0.64
6	0.82	0.64	0.66	0.71
7	0.82	0.57	0.46	0.62
8	0.79	0.50	0.39	0.56
9	0.93	0.71	0.69	0.78
Avg	0.87	0.66	0.61	0.71
StdErr	0.02	0.04	0.05	0.03

Table 2.2: Results of the generalization analysis performed on the study period data using the backpropagation classifier. The performance numbers presented in the table correspond to the proportion of correctly identified study period patterns. The average values are taken over all of the subjects, and the standard error is calculated over the subject means.

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.7622 (0.038)**	0.4386 (0.056)	0.4370 (0.055)
	Location	0.4972 (0.038)	**0.6258 (0.047)**	0.5007 (0.042)
	Object	0.4066 (0.061)	0.5123 (0.073)	**0.6697 (0.063)**

Table 2.3: Results of the backpropagation classifier applied to the recall-by-category periods of the experiment. As described in section 2.2.10, this table reports the average activity of each classifier output unit during verbal recalls from each of the three categories.

2.3.5 Analysis of final free recall with the classifier

The method of application of the backpropagation classifier to the final recall period data is described in section 2.2.10. The backpropagation classifier was trained only on data from the study periods of the experiment. The trained classifier is then tested on patterns from the final recall period. Figure 2.2 depicts the activation of the three outputs of the classifier over the entire course of the final free recall period, for a single subject. In order to compare these output traces to the behavioral data, we superimpose colored marks that represent the verbal recalls made by the subject.

The fluctuations of the classifier output correspond quite closely to the verbal recalls made by the subject. The classifier output seems to predict the behavioral performance of the subject, in terms of category of the item recalled. In the next section I quantify the correspondence between output trace and recall record.

Correlation and percent correct.

The correspondence between classifier output and verbal recall record is quantified in two ways: by calculating the pairwise correlations between each of the verbal recall records and each of the classifier traces and by calculating the percentage of recall events for which the classifier correctly predicts the category of the retrieved item (both described in section 2.2.11).

There are two effects to notice in the three-by-three correlation matrices (Table 2.6). First, the diagonal elements are positive, indicating that the rise and fall of the classifier output activities correspond to the verbal recalls made by the subject. Second, the off-diagonal elements are often negative, indicating that the classifier output activities tend to fall when a recall is made from a different category. I quantify this effect by taking the difference between the average of the on-diagonal elements and the off-diagonal elements (the OnOff metric, see section 2.2.11). The OnOff metric gives a sense of the overall ability of the classifier to track behavior (across categories).

It is difficult to assess the degree to which parametric statistical methods are appropriate to determine the significance of these classifier generated results. Thus, I calculated statistical significance using two non-parametric methods. The first method calculates within-subject significance

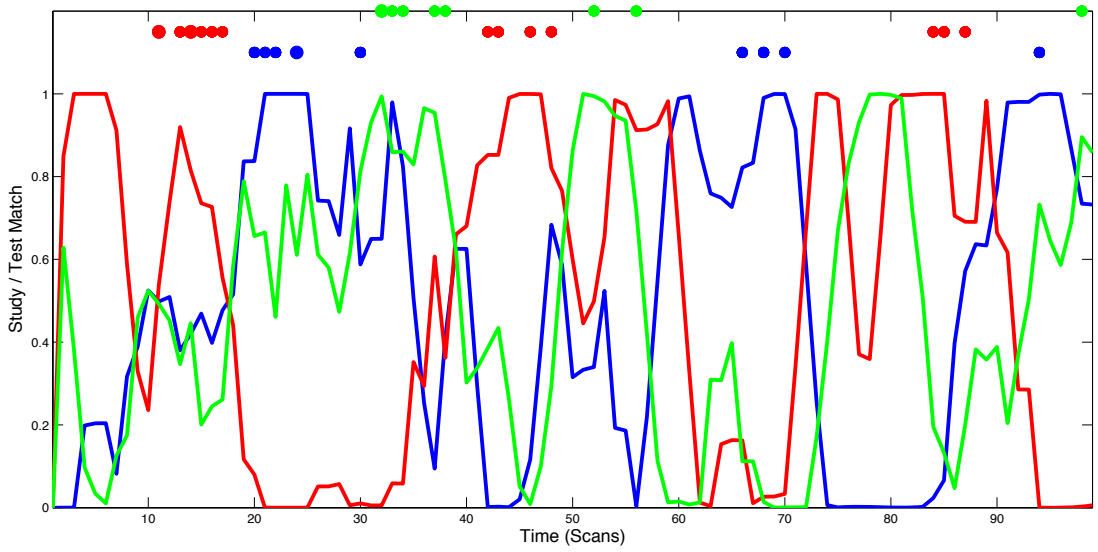


Figure 2.2: Time is represented on the x-axis; each tick represents one complete brain scan (see section 2.2.6). For each brain scan the classifier produces three estimates of match between the current testing pattern and the study categories. The blue, red, and green lines correspond to the face, location, and object output units, respectively. Activation values for the output units are represented on the y-axis. Above the graph of classifier output we plot the behavioral recall record for the subject. Here, the blue, red and green dots correspond to the face, location and object recalls made by the subject (larger dots correspond to multiple items recalled during a single scan). The recall events are shifted forward on the graph by three time-points, to account for lag to the peak hemodynamic response. For illustrative purposes, the classifier output lines are slightly temporally smoothed (see section 2.2.10).

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.3117**	-0.3401	0.0823
	Location	-0.1407	**0.2127**	-0.1480
	Object	0.0831	-0.1754	**0.3252**

Table 2.4: The correlation matrix for subject 9.

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.0100**	0.9997	0.2924
	Location	0.8020	**0.0529**	0.8133
	Object	0.2822	0.9429	**0.0061**

Table 2.5: The corresponding p-values for subject 9. These p-values were generated using a non-parametric statistical method, using surrogate data (see methods, non-parametric statistics).

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.3008**	-0.0945	-0.1341
	Location	-0.0629	**0.1415**	0.0288
	Object	0.0451	-0.0934	**0.1435**

Table 2.6: The average correlation matrix over all subjects.

of the correlations by generating surrogate versions of the classifier output with similar spectral characteristics (using a wavelet-based decomposition procedure, section 2.2.13). By doing this it is possible to assess the significance of both the individual elements of the correlation matrix as well as the OnOff metric described above. The second nonparametric method (also described in section 2.2.13) creates a distribution of surrogate OnOff values by scrambling the category labels on each subject's correlation matrix a large number of times. Using this surrogate distribution of OnOff values, it is possible to determine the probability that the observed results were obtained by chance.

I report significance values for the individual elements of a subject's correlation matrix (Table 2.5), a significance value for each subject's OnOff metric (Table 2.7, by subject), and a significance value for the entire experiment (Table 2.7, bottom). This final measure of significance does not rely on the validity of the surrogate data generated by the wavelet-based decomposition procedure. I also present a three-by-three correlation matrix averaged across all subjects (Table 2.6).

Subject	OnOff Metric	p-value
1	0.3844	<0.001
2	0.3008	<0.001
3	0.4233	<0.001
4	0.1721	0.032
5	0.2086	0.018
6	0.1569	0.025
7	0.0470	0.351
8	0.1412	0.066
9	0.3896	<0.001
	Mean OnOff	p-value
	0.2471	<0.001

Table 2.7: OnOff metric by subject with associated significance value. Below the by-subject data is the mean OnOff value over all subjects, with the associated significance value.

Subject	Percent Correct	p-value
1	0.7568	<0.001
2	0.5143	0.002
3	0.6512	<0.001
4	0.5172	0.018
5	0.5625	0.008
6	0.4167	0.081
7	0.4167	0.196
8	0.6154	0.001
9	0.8000	<0.001

Table 2.8: Percent correct by subject during the final free recall period. For a description of the statistical test used to generate the p-values, see section 2.2.13.

This correlation-based measure is not the only way to measure the correspondence between the classifier output and the verbal recalls. One can also measure, at the time of recall, whether the output unit with the greatest activity corresponds to the category of the item being recalled. By calculating the proportion of recall events that are correctly identified with this measure, I am able to generate a measure of percent correct. Table 2.8 reports this percent correct classification for each subject, and the associated significance values. The significance is calculated using the same surrogate data set (see section 2.2.13).

A few factors make this percent correct measure a sub-optimal means of measuring correspondence between classifier output and verbal recall. Rapid switches between categories and noise in the classifier output can conspire to reduce the overall percent correct, as our calculation does not allow partial credit when the correct output unit is second-most active. However, when consideration of this measure is paired with the correlation measure, one gets an adequate sense of the general ability of the classifier to track the subject's behavior.

2.3.6 Event related average.

Close visual inspection of Figure 2.2 as well as the traces generated for the other subjects suggests that a classifier output unit begins to increase its activity in anticipation of an upcoming verbal recall. To investigate whether this impression holds over the entire set of subjects, I constructed an event-related average of classifier output activity (see section 2.2.12). The events comprising this

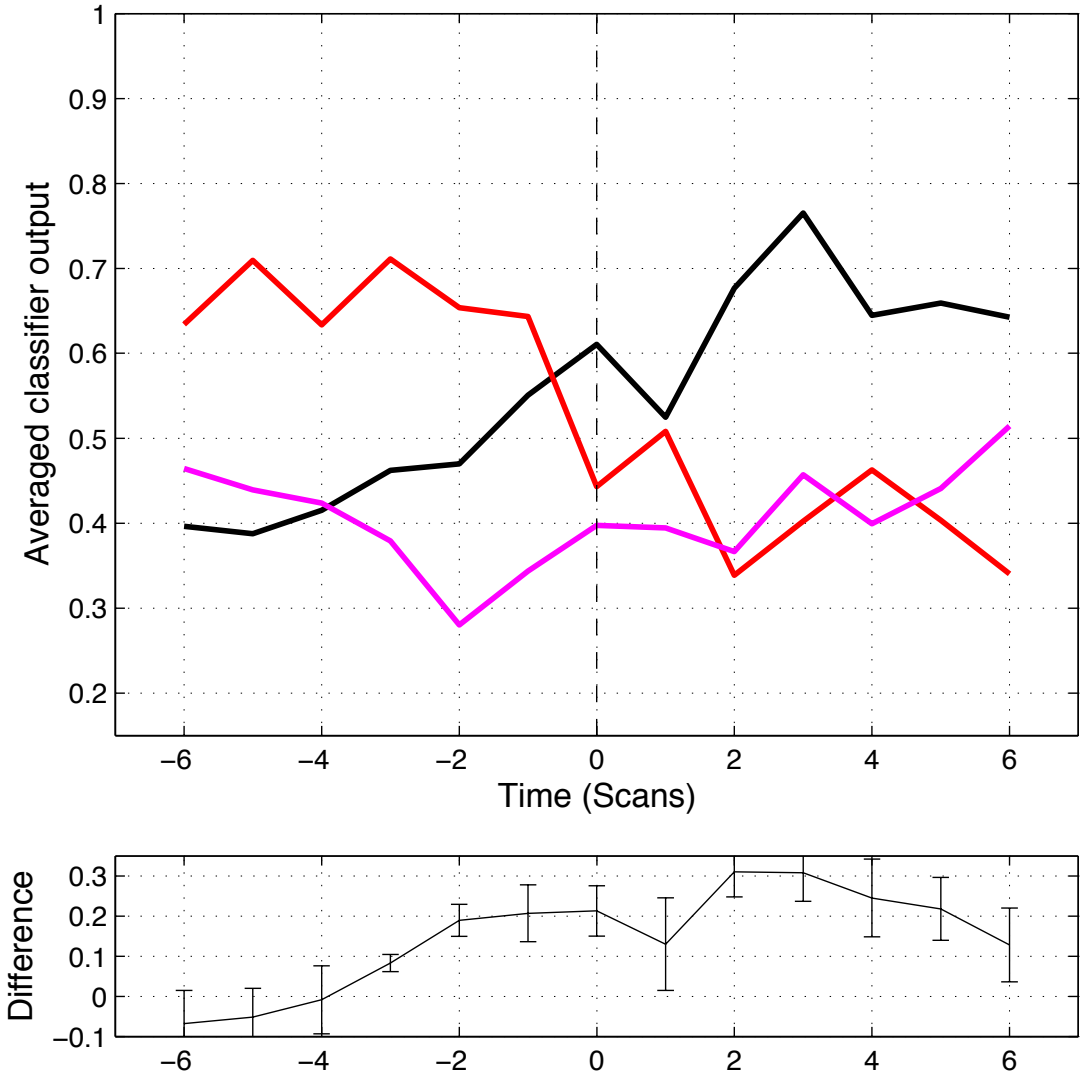


Figure 2.3: The top panel plots an event-related average of classifier output, over all subjects. The black line corresponds to the average output of the Currently Recalled category, while the red and purple lines correspond to the average output of the Recently Recalled and Not Recently Recalled categories, respectively (see text and section 2.2.12). The dotted line at x = 0 represents the scan on which the verbal recall was made; the line is not shifted to account for any hemodynamic lag effects. The bottom panel plots the difference between the Currently Recalled and Not Recently Recalled lines. See text for more details.

average were centered on a subset of correct verbal recalls; specifically, those recall events where there were no same-category recalls in an 8 scan window preceding the recall. This was done to ensure that the time-points in the graph that precede the recall event are uncontaminated by brain activity associated with previous recall-related events from the same category.

This event-related average collapses over category of item recalled. That is, the black line in Fig. 2.3 corresponds to whichever category is verbally recalled at the time-point marked by the vertical dashed line (the Currently Recalled category). Figure 2.3 shows two other traces as well, representing the activity in the other output units. I break the other output units into two categories: any category that has been recalled in the previous 8 scans is considered Recently Recalled (the red line), and any category that has not been recalled in the previous 8 scans is considered Not Recently Recalled (the purple line). The latter serves as a baseline or reference condition.

I am interested in whether the Currently Recalled category output trace rises significantly above the Not Recently Recalled category output trace before the recall is verbalized. I performed a pairwise t-test on the set of subject means that comprise each point, at each lag leading up to the recall event. The two lines significantly differ at time = -3 (that is, 3 scans or 5.4 seconds before the verbal recall), and on subsequent time-points (excluding time = +1). For this analysis the recall event onset was not shifted to account for hemodynamic lag. Thus, the neural event underlying this rise in category-related activity may take place earlier than this estimate.

2.3.7 Extracting brain maps from the classifier

The classifier was trained and tested on individual subject data. As such, the classifier was able to take advantage of idiosyncratic features of each subject's category response. In this analysis, I wish to inspect the patterns of activity that each classifier was sensitive to, in order to visualize the common features of the category responses across all subjects. The colormaps in Figure 2.4 represent the sets of voxels that are influential in determining the classifier output across all subjects. Section 2.2.10 provides details of the algorithm used to determine the relative influence of each voxel's signal on the activity of each category output unit are provided in section 2.2.10; this is termed the

"importance value" of each voxel, and is a function of the weight connecting the voxel to the output and the average activity of the voxel for that category during study (see equation 2.2). An individual map was generated for each subject; this map was converted to Talairach space (section 2.2.10) and averaged together with the maps from the other subjects, by category.

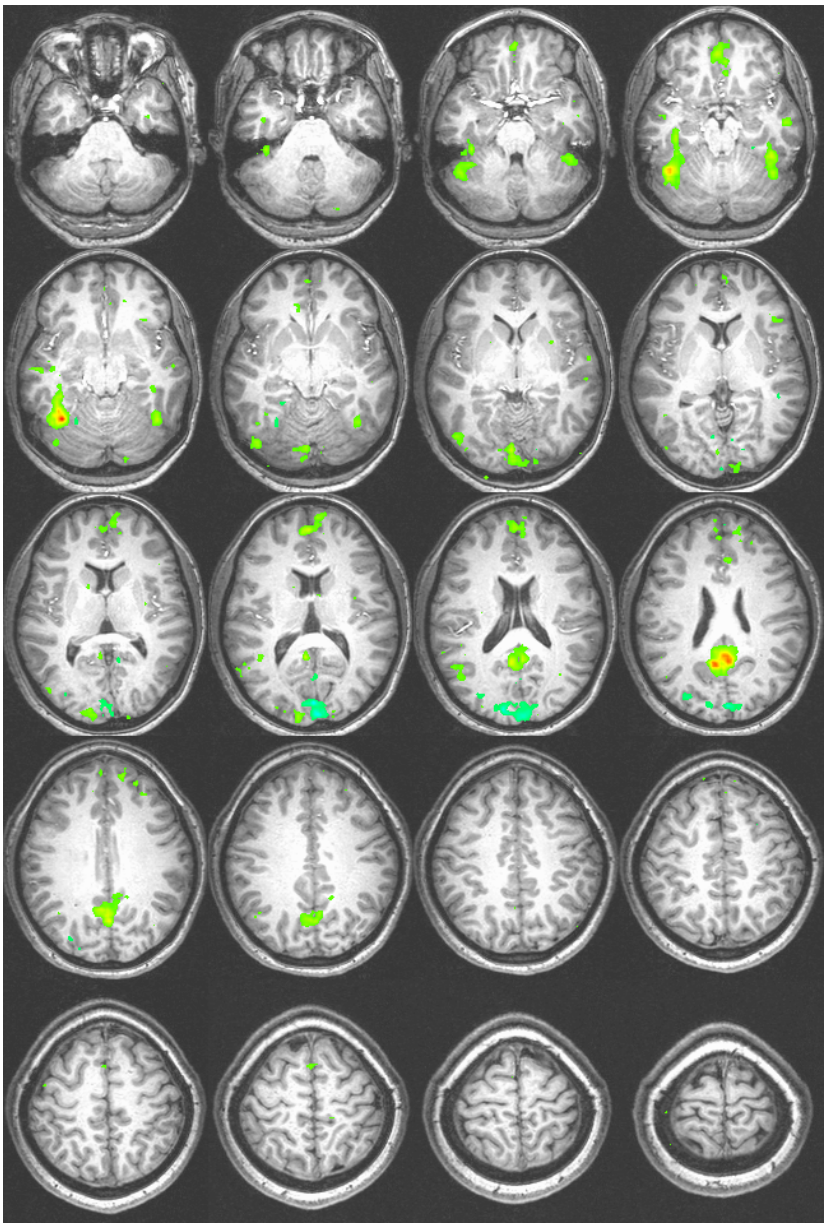
As one might expect, for the face and location tasks, areas in the fusiform gyrus and parahippocampal gyrus activate, respectively. Interestingly, even in this across-subject map, there seem to be a number of other areas that activate for each task. The classifier is able use the activity in these areas to generate predictions about subject behavior. There are two points to be kept in mind while considering these maps (as will be elaborated in section 2.4). First, these maps portray an across-subject response; in the previous analyses the classifier was run on individual subjects. Second, this type of analysis may over-emphasize canonical areas at the expense of areas that either activate less strongly or are idiosyncratic to a given subject, even though these other areas may be contributing to the classification (see section 2.3.9 for further exploration of this point).

2.3.8 Analysis of frontal and temporal lobe contributions

The brain maps presented in Figure 2.4 show sets of brain areas activated across-subjects for each of the three study categories. Interestingly, it looks like there is very little involvement of subregions of the frontal lobe (with the possible exception of the face map). Given our thorough discussion of the role of frontal lobe in free recall and memory targeting in Chapter 1, this finding deserves note. In order to determine whether this relative lack of frontal involvement is an artifact of the backpropagation classifier (for example, it is possible that given a strong temporal lobe signal, the classifier ignores signal in the frontal lobe).

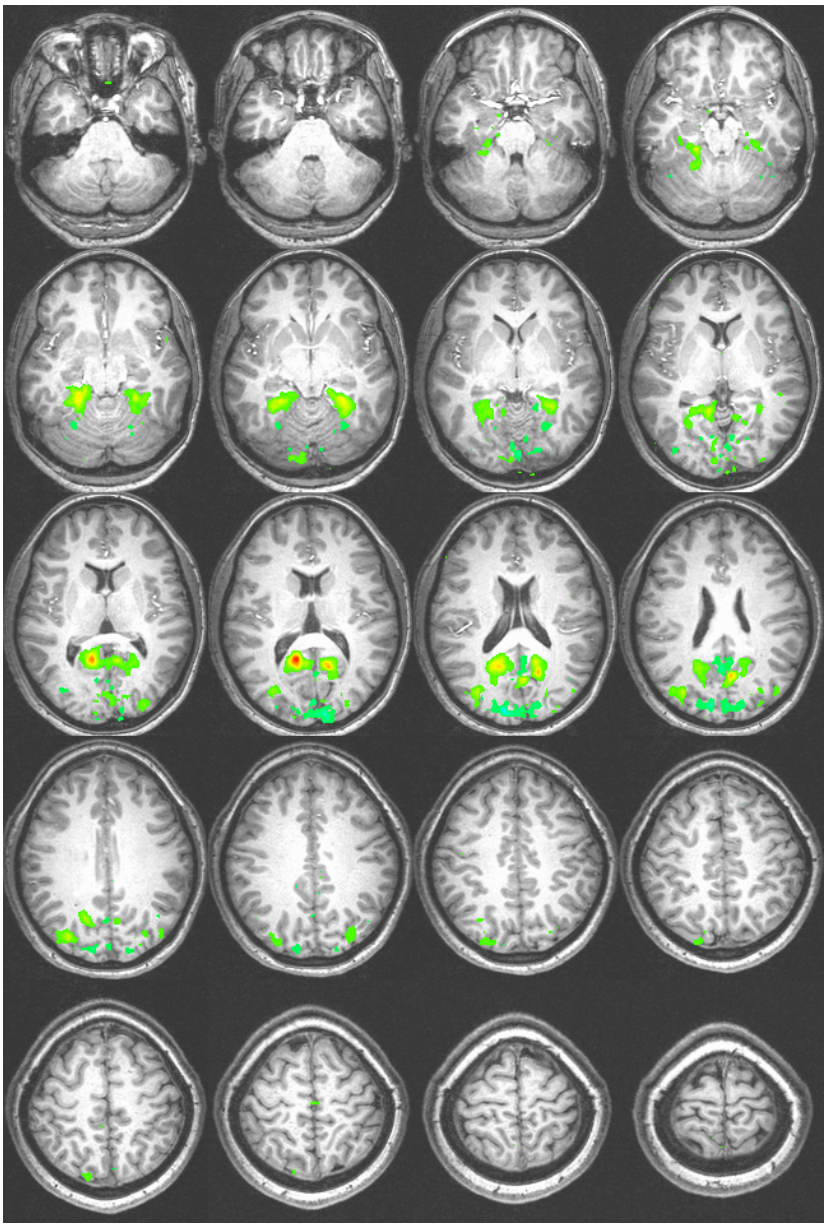
Section 2.2.8 describes the creation of two anatomical masks designed to investigate this issue. The first mask contains broad regions of the frontal lobes, and the second mask contains broad regions of the temporal lobes. The first set of results presented in this section was generated by training a backpropagation classifier on the frontal lobe data.

Table 2.9 contains the average correlation values (see section 2.2.11) for the backpropagation



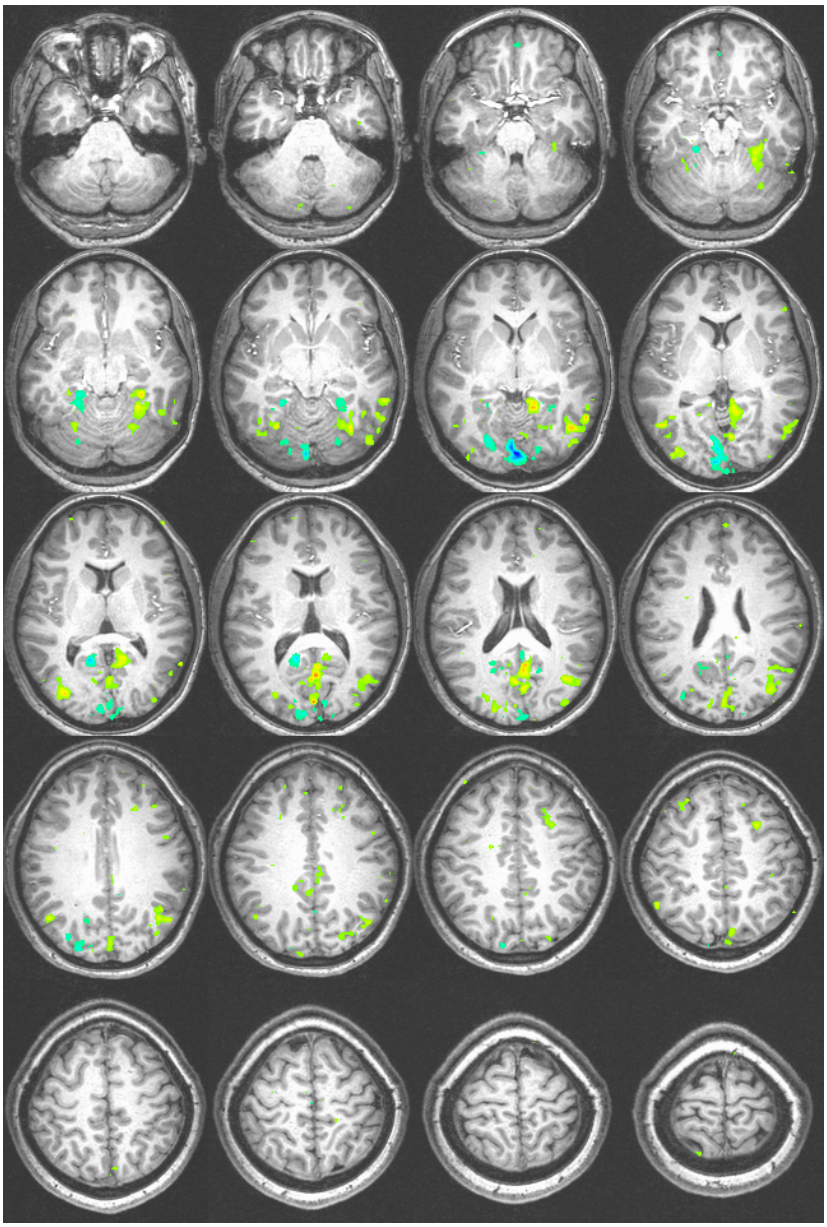
(a) The “study face” context

Figure 2.4: The classifier derived colormaps. See text for details.



(b) The “study location” context

Figure 2.4: The classifier derived colormaps (cont)



(c) The “study object” context

Figure 2.4: The classifier derived colormaps (cont)

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.160905**	0.009305	-0.152310
	Location	0.006503	**0.023692**	0.008824
	Object	-0.059629	0.014193	**0.052424**

Table 2.9: The average correlation matrix over all subjects, for the backpropagation classifier trained on voxels from the frontal lobe.

Subject	OnOff Metric	p-value
1	0.1854	0.0072
2	0.0209	0.4050
3	0.3195	0.0195
4	0.0297	0.3787
5	0.1011	0.1409
6	-0.0550	0.7557
7	0.1237	0.1119
8	0.0543	0.2522
9	0.1910	0.0131
	Mean OnOff	p-value
	0.1079	0.0034

Table 2.10: OnOff metric by subject with associated significance value, for the classifier trained on the set of voxels from the frontal lobe. Below the by-subject data is the mean OnOff value over all subjects, with the associated significance value.

classifier over all subjects, for the frontal lobe data. Immediately apparent is the weak correspondence between classifier output and subject behavior (for comparison see Table 2.6). Interestingly, the only cell of the correlation matrix that holds a large value is the face-to-face correlation. This is in line with the observation from the whole-brain classifier map (Fig. 2.4(a)) that the only category with a substantial representation in the frontal lobe was the face category. Table 2.10 shows the OnOff metric with associated significance for each subject. There is a wide variation of the classifier's correspondence to behavior across subjects. While three of the subjects have significant OnOff values, most are decidedly non-significant. The bottom panel of Table 2.10 shows that the results obtained over the entire set of subjects was significant (see section 2.2.13), but it is likely that this effect was carried by the face-related frontal pattern.

The second set of analyses presented in this section inspect the role of temporal lobe brain regions in identifying category-related brain activity during free recall. Table 2.11 contains the

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.308176**	-0.037670	-0.139475
	Location	-0.045295	**0.082080**	-0.050583
	Object	-0.040393	-0.050055	**0.134835**

Table 2.11: The average correlation matrix over all subjects, for the backpropagation classifier trained on voxels from the temporal lobe.

Subject	OnOff Metric	p-value
1	0.2349	<0.001
2	0.3795	<0.001
3	0.3339	0.0034
4	0.0602	0.2619
5	0.2131	0.0071
6	0.2286	<0.001
7	0.0854	0.1970
8	0.1708	0.0181
9	0.4143	<0.001
	Mean OnOff	p-value
	0.2356	<0.001

Table 2.12: OnOff metric by subject with associated significance value, for the classifier trained on the set of voxels from the temporal lobe. Below the by-subject data is the mean OnOff value over all subjects, with the associated significance value.

average correlation values (see section 2.2.11) for the backpropagation classifier over all subjects, for the temporal lobe data. Over all, the results correspond quite well to the corresponding whole-brain analysis (see Table 2.6). Interestingly, the location-location cell of the correlation matrix is reduced from the whole-brain analysis, suggesting that non-temporal lobe areas are contributing to the fluctuations of the location output unit in the whole-brain analysis. Table 2.12 shows the OnOff metric with associated significance for each subject. Unlike in the corresponding frontal lobe analysis, most of the subjects show a significant correlation of classifier to behavior, and the across-subject OnOff measure is significant.

It appears from these analyses that a substantial proportion of the category-related brain activity detected during the recall period arises from temporal lobe regions. In section 2.4.2 I discuss the reasons for this, both in terms of the situations in which we expect frontal lobe to be involved, and in terms of the mechanisms of the model described in chapter 1.

		Classifier Output		
		Face	Location	Object
Verbal Recalls	Face	**0.006716**	-0.147898	-0.041501
	Location	-0.059895	**0.030034**	-0.064301
	Object	-0.096992	-0.087827	**-0.037640**

Table 2.13: The average correlation matrix over all subjects for the ANOVA-contrast based average activity measure.

2.3.9 Analysis of average activity in ANOVA contrast masks

The analyses in this section were carried out in part to assess the advantage one gains in applying a non-linear pattern classification device (the backpropagation classifier) to neuroimaging data. Three category-specific average time-courses were constructed, as detailed in section 2.2.9. Each time-course consisted of signal averaged over a set of voxels identified by an ANOVA contrast as having a large category-specific effect during the study period. By conducting the same analyses on these time-courses as were conducted on the backpropagation classifier output traces, it is possible to assess the advantage gained by using the backpropagation classifier.

Table 2.13 contains the average correlation values, over all subjects, for the ANOVA contrast time series. In comparison to the correlations for the backpropagation network, presented in Table 2.6, these correlations are quite weak. Table 2.14 contains the OnOff metrics for each subject, with associated significance values. Only two subjects approach significance. Interestingly, the entire experiment achieves significance (Table 2.14, bottom panel), which seems counterintuitive, given the mediocre correspondence of average time-course to verbal recalls. However, close inspection of Table 2.13 shows that the on-diagonal elements of the correlation matrix are all slightly larger than the off-diagonal elements in the same row, suggesting that there is some weak correspondence of average activity to behavior. Inspection of the event-related average of the average time-course data shows that there is indeed a suggestive trend in this data. As seen in Figure 2.5 the black line (Currently Recalled) is consistently greater than the purple line (Not Recently Recalled). However, while this trend is in the right direction (black line greater than purple line), none of the differences reach significance.

Subject	OnOff Metric	p-value
1	0.1687	0.0306
2	0.0340	0.3497
3	0.1807	0.1317
4	0.0541	0.3021
5	0.1354	0.0814
6	0.0904	0.1671
7	0.0071	0.4733
8	0.0213	0.4087
9	0.0533	0.3001
	Mean OnOff	p-value
	0.0828	$<$0.001

Table 2.14: OnOff metric by subject with associated significance value for the ANOVA-contrast based average activity measure. Below the by-subject data is the mean OnOff value over all subjects, with the associated significance value.

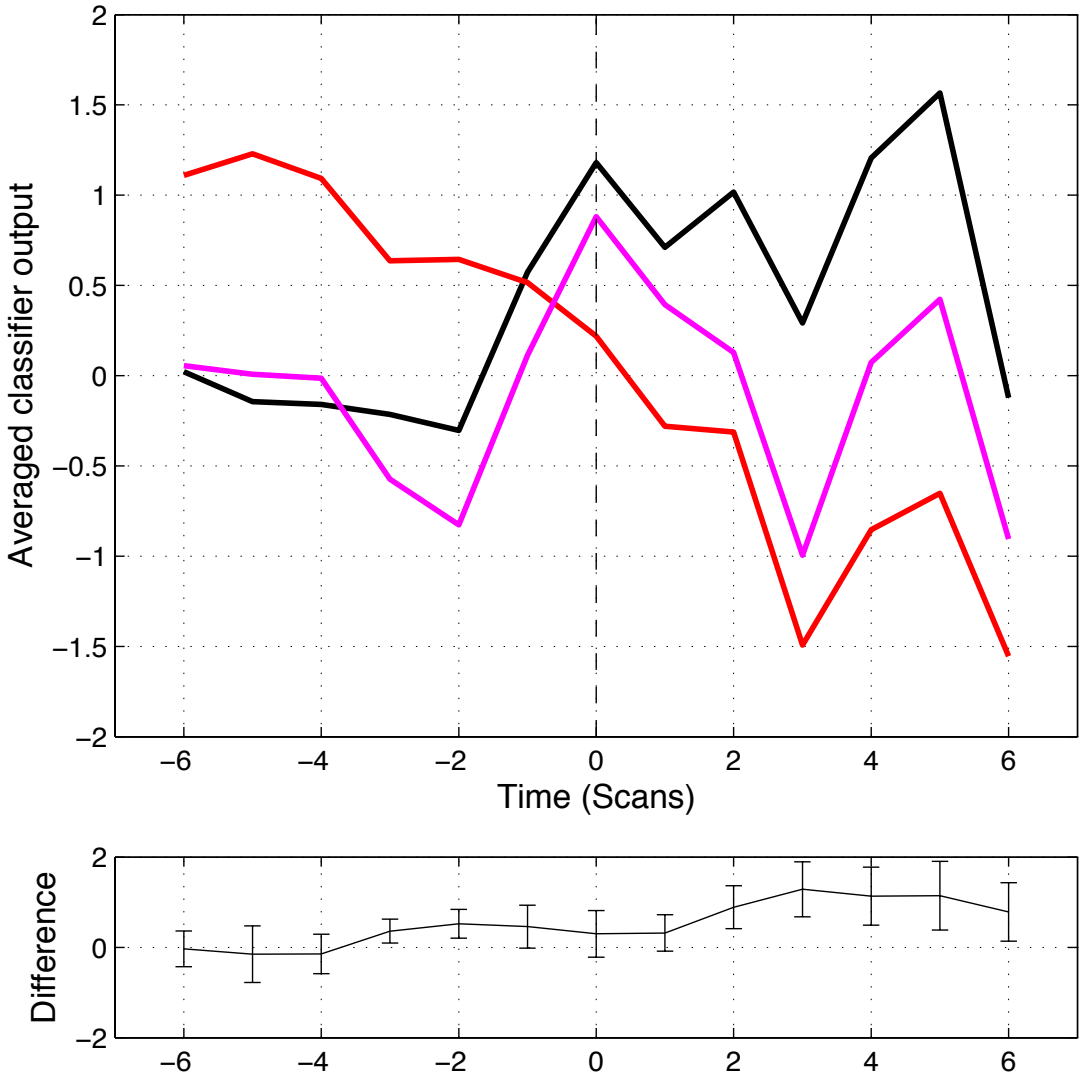


Figure 2.5: An event-related average constructed from ANOVA contrast voxel subsets, over all subjects. The black line corresponds to the average output of the Currently Recalled category, while the red and purple lines correspond to the average output of the Recently Recalled and Not Recently Recalled categories, respectively (see text and Methods, event-related average). The dotted line at x = 0 represents the scan on which the verbal recall was made; the line is not shifted to account for any hemodynamic lag effects. The lower graph corresponds to the difference between the Currently Recalled and Not Recently Recalled lines. See text for more details.

2.4 Discussion

In this chapter I describe results from an fMRI study of free recall. A backpropagation classifier is used to identify patterns of brain activity related to the study of three categories of items. The patterns of brain activity detected by the classifier seem to be relatively consistent across items and runs (section 2.3.3). Furthermore, the strength of each category pattern during the recall period showed a significant correlation with the verbal recalls made by the subject (section 2.3.5). Inspection of an event-related average of the classifier output showed that the classifier's detection of a category-related pattern tended to precede the subject's response by a few seconds (section 2.3.6), consistent with the idea that the detected patterns were being used to cue the memory system.

Subsequent analyses showed that areas across the brain were responsive to each of the studied categories (section 2.3.7). Reanalysis of anatomically scoped subsets of the original data suggested that temporal lobe regions were more important to the classification than frontal regions (section 2.3.8). A reanalysis using averaged activity of category-sensitive voxel subsets suggested that the feature-weighting and nonlinearity of the backpropagation classifier were important factors in detecting the recall-related patterns (section 2.3.9).

2.4.1 Interpretation of the detected patterns

In section 1.2.4 I enumerated a set of criteria that a pattern of brain activity would have to fulfill to be considered related to memory targeting. The pattern of activity would have to be present during encoding, present during retrieval attempt, and its strength should correlate with recall performance. It is interesting to note that the pattern of brain activity corresponding to the item representation itself would show a similar profile. Several researchers have noted this difficulty in in distinguishing between cue-related and retrieval-related activity fluctuations (Wheeler & Buckner, 2003; Kahn et al., 2004).

From a theoretical standpoint, this distinction may be a false dichotomy. "Cue-related" refers to some representation that is being projected to the memory system in order to retrieve stored infor-

mation: the "retrieval-related" representation. However, in this experiment, a reasonable retrieval strategy for a subject to use would be to think broadly about the categories presented during the study period. By reactivating a cortical pattern broadly consistent with "face", the memory system would be biased to retrieve memories of the studied celebrities. Given a successful retrieval, face-related features would be reactivated in the very same cortical areas used to cue the memory system. Thus, by this framework, the same areas that were used to cue the memory system would also be the recipients of the retrieved content. To the extent that one can make their brain "look like" it did at study (by thinking broadly about the types of items studied, by thinking about the task performed at study, etc.), one will be more successful in retrieving items.

The theory predicts that cue-related and retrieval-related patterns of activity can be supported by the same brain regions. However, in future studies it may be possible to design an experiment where the memory system is cued with one type of information and retrieves a second. For example, one can imagine a source memory study in which celebrities are paired with either locations or objects at study, and at test are given the celebrity name as a cue. This type of dissociation might allow us to get a better sense of the time-course of each process.

2.4.2 Brain maps and anatomical considerations

In section 2.3.7 I present a series of importance maps in which each voxel is colored in proportion to its influence on the classifier output unit corresponding to a given category. The role of PFC in memory targeting receives a great deal of attention in chapters 1 and 4, but robust activity patterns in PFC are conspicuously absent from these maps (shown in Figure 2.4). Inspection of the targeted anatomical analyses (section 2.3.8) suggests that in this experiment PFC areas are only reliably engaged by stimuli from the face category.

As mentioned in section 2.1, the three stimulus categories were chosen because they are known to evoke distinct activity patterns in the ventral temporal lobe (Haxby et al., 2001). By the argument posed in the previous section (2.4.1), a reasonable strategy for a subject to undertake to retrieve stored memories would be to cue the memory system with activity patterns broadly consistent with

those corresponding to the studied categories. Thus, in this experiment, it is reasonable to expect that the memory targeting is carried out by temporal areas.

There are a few types of memory-related brain activity that will not show up in the current analysis. These hypothetical patterns of brain activity do not fit the profile of a memory targeting representation. For example, it is possible that areas of PFC are differentially activated by the three category types during the recall period, but are not reliably activated during the study period. In the current analysis, the classifier was trained on study period data. Thus, if a brain region was not activated during the study period, it will not show up on the importance map.

It should be possible to investigate the hypothesis that certain patterns of retrieval-related brain activity are present at recall but not at study. By training a classifier on brain data from the recall-by-category periods (section 2.2.4) and testing on data from the final free recall period, areas of this sort should be detected. The maps derived from this analysis can be compared to the maps in Figure 2.4, to see if any additional brain regions appear.

Furthermore, any brain activity that does not differentiate between the three stimulus categories will not be detected in the current analysis. A number of brain regions (including several in PFC) reliably activate during memory retrieval (Wheeler & Buckner, 2003; Kahn et al., 2004). However, as mentioned, the classifier was designed to detect brain patterns that correlate with the three stimulus categories, and a follow-up analysis will have to be done to characterize the other types of activity patterns often seen in fMRI studies of memory retrieval.

A similar argument explains why I did not detect activation in the hippocampus. Hippocampus ought to be reliably activated during study and retrieval regardless of stimulus category. Thus, I do not expect to see differential hippocampal activity in the current study; the classifier will only detect areas whose activity patterns are distinct for the categories.

A final point about the brain maps regards the presence of negative importance values for a given category. Backpropagation, as an error-driven algorithm, learns both that a given pattern belongs to the face category, and that it is not a member of the location or object category. Thus, it will tend

to pit the categories against one another; evidence used to activate one output unit can be used to deactivate another. This may partially explain the negative correlations in the off-diagonal cells of the correlation matrices presented in Tables 2.4 and 2.6.

If these negative weights were quite strong, then it is possible that the activation of a given category output unit could be due to the weakening of another category's pattern. However, if this was the case, one would not expect this output unit's rise to predict the identity of the next recalled category. In Figure 2.3, the black (Currently Recalled) and the purple (Not Recently Recalled) lines both represent categories that have not been recalled recently. If the classifier was not sensitive to the upcoming category identity, one would expect both the black and purple lines to rise. However, the classifier seems to have information regarding the identity of the Currently Recalled category, and the black line rises above the purple line before the recall is verbalized.

2.4.3 Comparison of backpropagation to ANOVA contrast averages

In section 2.3.9 I describe results of an analysis in which I created three category-selective masks, each consisting of a set of voxels identified by an ANOVA contrast to be selectively responsive to one of the categories (see section 2.2.9 for more details).

By averaging the activity of each set of voxels, I created three time-courses, one for each of the three stimulus categories. I ran the same set of analyses as was run on the output of the backpropagation classifier (section 2.3.5) on these time-courses (correlations and event-related average). It is clear, from comparing the results, that the backpropagation classifier is much more sensitive to the patterns of brain activity that appear during the recall period.

There are a number of features of the backpropagation classifier algorithm that may confer this advantage. The backpropagation classifier weights the features of the input space, and passes the weighted sum through a non-linear (sigmoid) transfer function to arrive at an estimate of category strength. In the ANOVA contrast, the features are simply averaged together. While the ANOVA contrast time-courses were not as well correlated with the behavioral data, it is clear from inspecting the results of the event-related average (Fig. 2.5) that there was a trend toward correspondence.

2.4.4 Conclusion

There is much work yet to be done to unravel the processes by which humans mentally voyage into their own past. The current study offers a first pass characterization of the elaborate mechanisms that drive memory search and retrieval. Part of the value of this endeavor is proof-of-concept. At the outset it was unclear whether the methodological hurdles of running a free recall experiment in the scanner could be overcome, and if overcome, whether there would be any interesting results to report. However, we now have a paradigm that produces interesting data in the domain of free recall. This paradigm can now be modified to allow us to investigate more elaborate hypotheses regarding the nature and possible dissociation of cue-related and retrieval related processes in the human brain. Some potential future directions are outlined in chapter 5.

Chapter 3

The effect of task on memory accessibility in free recall

3.1 Introduction

In section 1.2.1, I reviewed a number of types of context representations that affect memory accessibility, including environmental context, temporal context and task context. Specifically, I reviewed evidence that a shift in context during the study period (such as by changing rooms in the case of environmental context) will have an effect on the accessibility of memories encoded before the context shift. These context-related effects can provide constraining data for models of human memory. As such, it is worth considering the behavioral effects of different types of context change.

In the literature, there are few studies of the effect of encoding task context on memory accessibility. The studies that do exist report only serial position effects, which can obscure interesting behavioral effects seen in output transition probabilities (Kahana, 1996). As mentioned (section 1.2.1), Watkins and Peynircioğlu (1983) showed that subjects can successively target items studied with three distinct encoding tasks, and show a strong recency effect for each item. However, in this study subjects recalled the items associated with each task separately, so it is not possible to assess the degree to which items encoded in the context of each task were isolated from each other in mem-

ory. There have been a few studies (Koppenaal & Glanzer, 1990; Thapar & Greene, 1993; Neath, 1993) investigating the effect of distractor task switch on memory accessibility. I will return to the relationship between encoding task and distractor task in chapter 5.

In this chapter, I interpret the behavioral effects of an encoding task switch in terms of a context-cued memory system. I describe a continuous distractor free recall experiment in which I attempt to systematically manipulate the subject's inner mental context by switching between two encoding tasks, a given item is either studied using one or the other. The first task is a pleasantness judgment: subjects characterize an item as good or bad. The second task is a size judgment: subjects classify an item as bigger or smaller than a shoebox (see section 3.2.4). With this design, I establish two states that the inner mental context of a subject can be in, the pleasantness state, and the size state. I will use the phrase "task state" to refer to the subset of inner mental context that varies predictably with task. By associating subsets of items on the same study list with each of the two tasks, I hope to segregate the two sets of items in memory, and then show the effects of this segregation behaviorally. In the current study encoding task either switches halfway through a study list, or remains the same throughout the list.

I hypothesize that task state forms a large part of the cue used to search through memory, even when the subject is not performing the task. In terms of the model, a task-related pattern is still active in the PFC component as recall begins. This pattern is projected into the memory system, and the traces most likely to be recalled correspond to items studied with that task. Given a context system that can update during recall (see sections 1.2.2 and 1.3), the subject should still be able to access items from the first half of the list. However, given that each set of items is encoded with a different context, this should reduce the probability that items studied with different encoding tasks will be recalled successively.

3.2 Methods

3.2.1 Overview

The experiment was designed and presented using E-Prime 1.1 (Psychological Software Tools, Pittsburgh, PA). Subjects read a series of slides detailing the experiment and were then asked to repeat back the basics to the experimenter, who filled in any gaps in understanding. Subjects then performed the four practice lists, one of each type (see below). This was followed by the eight experimental lists. The subject was then debriefed.

3.2.2 Subjects

22 Undergraduates at Princeton University (15 female) participated in this experiment for course credit and payment.

3.2.3 Materials

A set of Matlab scripts were developed to automate the creation of randomized word lists. The larger wordpool used was a subset of the Toronto Noun Pool (acquired from the Kahana lab web site: `http://memory.psych.upenn.edu`); the subset was picked by hand, to exclude words inappropriate to the current encoding tasks. This was done to exclude words too abstract for the encoding tasks (section 3.2.4), for example, the word "absence" was excluded. The wordpool included the Kučera-Francis frequency (Kučera & Francis, 1967) of each word, as well as a matrix of word similarities, drawn from the LSA (Landauer & Dumais, 1997) database. These measures were used to automate list creation, as described below.

The words from each list were chosen randomly without replacement from the larger pool. The set of words was then subject to a series of tests, to determine if it was a suitable list. If all of the lists in a set passed the tests, those 12 lists were presented to a subject in the experiment. Lists were tested to ensure that the mean frequency value of the list fell within a certain range (20 to 50), and that the frequency variance fell within another range (100 to 6000). These numbers were arrived

at by generating a large number of sample lists, and inspecting histograms of mean and variance of frequency across randomly generated lists. The threshold values above were chosen to include the modes of the distributions, and exclude the long tails. This was done to minimize sources of behavioral variance attributable to presentation of lists with wildly different frequency profiles.

Two more tests were performed to determine list suitability using the matrix of LSA word-similarities. A threshold was set (0.25) to exclude lists where any pair of words had a similarity value that exceeded this threshold. Another threshold was set (0.70); if any two words in the entire set of 12 lists had a similarity that exceeded this number, the entire set was thrown out. This restriction procedure was applied to minimize semantic similarity effects on transitions (Howard & Kahana, 2002b).

3.2.4 Behavioral procedure

The first four study / recall blocks were considered practice, for the purpose of familiarizing subjects with the tasks. Every subject had the same task order for the first four lists: Pleasantness Control, Size Control, Pleasantness-Size Halfway, Size-Pleasantness Halfway. These blocks were not included in any of the present analysis. The remaining eight lists were assigned task in a pseudo-random fashion, each of the four variants appeared in each set of four lists, but were randomized within that set.

The study period

Subjects studied a list of 12 words for a subsequent free recall test. There were two encoding tasks used in this experiment (encoding task A, rate pleasantness of the item; encoding task B, judge whether the item is bigger or smaller than a shoebox). Each task had two labeled keys associated with it ("good" and "bad"; "big" and "small").

Each study trial was composed of a cue slide (lasting 1.5 sec), a stimulus slide (lasting 1.5 sec) and a blank gap (lasting 0.5 sec). Each study trial was followed by a distractor period (lasting 9 sec), described below. The cue slide consisted of a word orienting the subject to the upcoming encoding

task (“pleasantness” or “size”). The stimulus slide added the study item below the task word. The study item was presented in capital letters. A tone sounded and a warning message was displayed if the item was encoded with the wrong task (if the wrong key was pressed) or if the subject was slow to respond.

Study trials were embedded in one of two study conditions: *halfway*, in which the first half of the items were encoded using task A and the other half were encoded using task B (and vice-versa; the switch took place between the sixth and seventh items); and *control*, in which all items were encoded with a single task.

A distractor task (counting backward by sevens from a three-digit number) preceded the first item and followed every item. The primary purpose of the distractor task was to disrupt rehearsal and to prevent the subject from forming associations directly between items. Every distractor period lasted for 9 seconds (1.5 sec title slide, 7.5 sec counting backward).

An experimenter was present in the testing room for the duration of the experiment. In the event that a subject did not count backwards, or did not perform appropriately during the experiment, they were reminded by the experimenter of the appropriate protocol.

The recall period

After the final distractor period, there was a cue to begin recall of the most recently presented list (a tone and three asterisks). The recall period lasted for 50 seconds. Subjects were told to report as many items as they could remember from the most recent study list, in any order, and were instructed to attempt to recall items through the entire period.

3.2.5 Behavioral analysis

Recording and scoring verbal responses

Verbal responses were recorded using a MicFlex USB microphone (MacMice, USA) connected to a Powerbook G4 Apple computer running Audacity (open source recording software: `http://audacity.sourceforge.net`, see section 2.2.4) and parsed with software devel-

oped by the Kahana lab (`http://memory.psych.upenn.edu`). Incidental verbal utterances as well as intrusions were excluded from the current analyses.

A recall was classified as valid if the item recalled came from the current list. Repeated recalls of items were counted as valid recalls, unless they were immediately repeated. Items from previous lists, or from extra-experimental sources (intrusions) were coded as invalid, and were not included in the current analyses. A given transition between items during the recall was considered valid if it was between two valid recalls.

Statistics

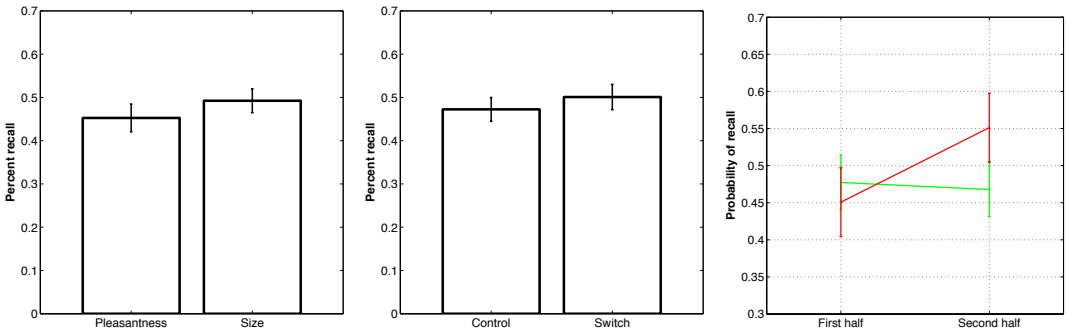
Unless otherwise mentioned, all error-bars represent standard error on individual subject means for the given measure. The two by two ANOVA was run using the JMP-IN software package on MacOSX. Subjects were treated as a random effect. All other statistics were performed as t-tests on subject means.

3.3 Results

Over the course of the experiment subjects performed two types of encoding tasks (section 3.2.4). The two tasks were designed to evoke similar depth of processing. As shown in figure 3.1(a) subjects performed approximately equivalently in the two tasks (paired t-test on mean percent correct by task, $p>0.1$). For the rest of the analyses I collapsed over task, as performance seemed to be comparable.

A two by two ANOVA was run (list type x list half; using the JMP software package) on percentage of items recalled. Subjects were treated as a random effect. As depicted graphically in figure 3.1(b) there was no main effect of list type (control versus switch; $F(1,21)=2.51$, $p>0.1$). Nor was there a main effect of list half, collapsing over list type ($F(1,21)=1.70$, $p>0.2$). However, there was a significant interaction of list type and list half ($F(1,21)=5.94$, $p<0.05$).

Figure 3.1(c) shows performance broken down by list half and list type. I performed two contrasts to characterize this interaction between type and half. First, I compared first half recall to



(a) Percent recall by task

(b) Percent recall by list type

(c) Probability of recall by list type and list half. The green line corresponds to the control condition and the red line corresponds to the switch condition.

Figure 3.1: Probability of recall over the entire list, broken down by list type, task performed and list half. See text for associated statistics.

second half recall in the control condition and in the switch condition. This contrast was not significant in the control condition ($t(21)=0.297$, $p>0.2$), but was significant in the switch condition ($t(21)=-3.15$, $p<0.005$). In other words, subjects performing control lists recalled an equivalent number of words in the first and the second halves of the list, but recalled significantly more words in the second half of a switch list, when compared to the first half of the switch list.

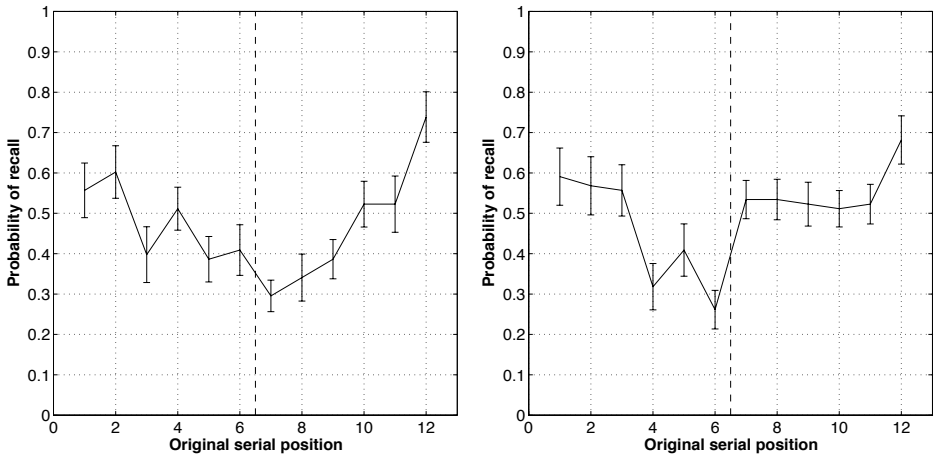
Furthermore, percent recall for the first half of the lists was not significantly different by list type ($t(21)=0.832$, $p>0.2$). However, subjects did recall significantly more items in the second half of a switch list compared to a control list ($t(21)=-2.61$, $p<0.05$).

In order to more thoroughly investigate the pattern of results obtained here, I broke down percent recall by serial position in the study list. Figure 3.2 shows the serial position curves by condition. The most salient feature of the serial position curve corresponding to the switch condition (Fig. 3.2(b)) is the large discontinuity between the sixth and seventh items, corresponding to the time that the subjects switch encoding task.

The theory predicts that there will be a large shift in internal context at the time of the task switch. It follows that the item immediately preceding the switch will be most harmed by this shift in context; the likelihood of a backwards transition landing on this item should be significantly reduced. I quantify this not by looking at the transition probabilities, but rather by investigating the simpler measure of probability of recall by serial position. Indeed, probability of recall for the item immediately preceding the switch is significantly decreased (as measured by a paired t-test; $p<0.05$).

Thus, there is a benefit of task switch, in terms of recall probabilities, for the second half of the list, as well as a cost of task switch, to the item preceding the switch. One might have expected the entire first half to show a cost, and while there is a trend in this direction, it is not significant. I will revisit this issue in the discussion.

Beyond these measures of percent recall, the model makes more nuanced predictions about the behavioral differences between the control and switch conditions, regarding the types of transitions



(a) The control condition

(b) The halfway switch condition

Figure 3.2: Probability of recall by serial position at study. The second half of the study list shows a significant increase in recall probability in the switch condition (see text on ANOVA for statistics on list halves). Furthermore, the sixth item in the switch condition shows a significantly reduced probability of recall (see text for statistics).

	Cross-half transition probability	Standard error
Control	0.2791	0.018
Switch	0.2341	0.015

Table 3.1: Probability of an output transition that crosses the list half boundary, broken down by list type. Standard errors are calculated across subject means. The difference between conditions is significant at $p < 0.02$ (One-tailed paired sample t-test).

that subjects make on successive verbal recalls. Broadly, the theory predicts that while recalling words studied with one encoding task, words studied with the other encoding task will be less accessible.

Specifically, I predicted that the subject would be less likely to transition across list half in the switch condition. Table 3.1 shows the probability of the subject making an output transition that crosses the list half. To arrive at this number, the number of cross-half transitions was divided by the total number of transitions.

As shown in table 3.1 there is a significant difference in the probability of a cross-half transition for control versus halfway switch (as determined by a one-tailed paired sample t-test on the cross-

half transition probabilities, $p<0.02$).

3.4 Discussion

As might be expected, an encoding task switch seems to reduce the proactive interference on the second half items, which is consistent with the idea that the items associated with each task are segregated in memory. Furthermore, given a task switch, there is a significantly reduced probability of transitioning across the list half during recall. Not only do subjects recall more second half items in the switch condition, they seem to be stuck in this portion of the list. These data points are discussed further below.

In chapter 4 I describe simulations of the effects of task switch, using a computational model of the human memory system. The model captures the reduced probability of cross-half transition and the enhanced probability of recall for post-switch items. However, the model is also significantly impaired at recall for the pre-switch items, which is not consistent with the current behavioral data (Figure 3.1(c)). The behavioral data show that the proportion of first half items recalled is not significantly reduced by a task switch.

The pattern of results produced by the model is more in line with an environmental context change, or a directed forgetting paradigm, in which there are both costs and benefits to the context change (as opposed to just benefits, as seen here). In section 4.4.2, I suggest a modification to the model that could allow it to better capture this pattern of results.

3.4.1 Task and memory accessibility

During the encoding period, task is an important part of inner mental context. This follows from the idea that task performance is guided by tonic activity patterns in frontal regions of the brain. Any tonic pattern that projects to the memory system becomes an effective context, and is embedded in any memory traces formed while it is present. Thus, the list-halves become isolated from each other in memory due to the fact that the task component of context has changed between halves.

If this characterization is correct, I expect that during retrieval one can characterize the internal contextual state of the subject by looking at which items are being retrieved. Given a recall of an item associated with Task A, items associated with Task B should be less accessible.

Our measure of transition probability quantifies and validates this prediction. While subjects in the switch condition make more transitions overall, fewer of these transitions cross the encoding task boundary. In other words, in the switch condition, subjects are less likely to transition between halves.

The phenomenon of increased second-half recall (Fig. 3.1(c)) can be described as a reduction in proactive interference from the items in the first half of the list. The model describes the mechanism by which interference is reduced; the cue being used to probe memory is a better match for the second half items than the first half items.

Chapter 4

A computational model of the memory system

4.1 Introduction

In the opening chapter of this dissertation, I described a computational model of memory search at a schematic level, focusing mainly on how it informs each section of the thesis. Here, I undertake a detailed description of the model, to explore the set of mechanisms necessary for its proper function, as well as to serve as a recipe of sorts, to guide future simulation projects. Specifically, I review the principles driving the function of the components of the model, highlighting the core mechanisms responsible for the phenomena discussed in the other chapters. A series of simulations are performed with the model, as described below.

In the first section (4.3.2) I simulate the basic free recall paradigm. The model exhibits the lag-recency effect (see section 1.2.1; Kahana, 1996) and captures the shape of PFR and CRP curves seen in behavioral studies of free recall (e.g. Fig. 1.1 and Fig. 1.2). In section 4.4.1 I describe the relationship of the current model to the TCM model, and describe potential modifications to the model, to increase fit to behavioral results.

Chapter 3 of this dissertation describes a study that manipulated encoding task identity during a

free recall paradigm. The results of this study were interpreted in terms of task representations and their effect on the accessibility of sets of memories. In this experiment, subjects studied half of a list using one encoding task and the other half using a different encoding task (see section 3.2 for more details). In this halfway switch condition, subjects showed a reduced probability of making an output transition that crossed this list-half boundary, relative to the control condition. Furthermore, subjects on average recalled more post-switch items in the halfway switch condition (relative to the same serial positions in the control condition; section 3.3).

In section 4.3.3 of this chapter, I describe results of a set of simulations of these encoding task effects during free recall. In order to simulate the effects of task switch during the study period, I created two orthogonal task representations, each corresponding to one of the two tasks used in the behavioral experiment. In the halfway switch condition, one representation was activated in the PFC component of the model during the first half of the study list, and the other representation was activated during the second half of the study list (see section 4.2.7 for more details). The basic results of the behavioral study are captured in a qualitative way by the model (section 4.3.3). Potential modifications to the model, to increase fit to the behavioral results, are described in section 4.4.2.

In section 1.2.3, I described a set of findings related to the characterization of the cognitive system and the healthy aging process. A study of free recall in young and older subjects (Kahana et al., 2002) showed that older subjects exhibit a flattening of their conditional response probability (CRP) profiles, but spared probability of first recall (PFR) profiles as shown in Figures 1.1 and 1.2. In section 4.3.4, I describe results of a set of simulations that attempt to fit this profile of deficits. I subject the model to a wide array of types of damage, and find that the only damage that adequately explains the pattern of results in the free recall study is damage to the gating system of the model. In section 4.4.3, I describe how this explanation differs from the explanation presented in Howard et al. (submitted). I also describe how a study of elderly subjects in a continuous performance task (AX-CPT;Braver et al., 2001) may provide convergent evidence for this type of damage to the gating system in older people.

4.2 Methods

4.2.1 Simulation package

All simulations were run with the PDP++ software package, which is available at `http://psych.colorado.edu/~oreilly/PDP++/PDP++.html`. Within the PDP++ software, I used the leabra++ package, which provides customized algorithms for running connectionist simulations. A detailed description of these algorithms is available in O'Reilly and Munakata (2000). The simulations described in this chapter used a slightly modified version of the leabra++ software (sleabra++), including a modification to the inhibition algorithm (section 4.2.2) and a modification to the PFC component (section 4.2.5).

4.2.2 Algorithm details

The leabra algorithm

The leabra algorithm incorporates six principles for development of connectionist models of cognition, as follow: biological realism, distributed representations, inhibitory competition, bidirectional activation propagation, error-driven task learning and Hebbian model learning (O'Reilly, 1998, 2001). The leabra++ software implements these principles using a point-neuron activation function, a k-Winners-take-all inhibition algorithm, and error-driven and Hebbian weight change equations. These are reviewed below.

The processing hierarchy

The leabra algorithm breaks processing up into trials, phases, and cycles. In the current simulations, each trial corresponds to the presentation of an item during the study period, and a recall attempt during the recall period (section 4.2.6). Each trial consists of three phases, and is implemented using the 'MINUS PLUS PLUS' phase order in leabra++. The three phases are required for calculating updating in the gated prefrontal component (section 4.2.5). Each phase consists of a number of cycles, during each of which the excitatory input to all the cells is calculated, the KWTA

inhibition is calculated for each layer, and the activation value for each unit is calculated. A given phase is ended if either of two criteria is met: a maximum number of cycles is reached, or the average change in activation (da) across cycles falls below some critical value (0.005). The maximum number of cycles in a phase during the study period was 1800, and during the recall period was 2000.

Bias

As mentioned in section 1.3, several models of recall contain mechanisms to prevent the network from persistently recalling the same memories (Raaijmakers & Shiffrin, 1981; Becker & Lim, 2003). I implemented a simple "tiring" mechanism, by which a given unit's bias weight is decremented in proportion to its final activity level on a given trial of the simulation. The bias weight determines the strength of a tonic input to the unit. By decrementing the bias weight the units is made less active on the next trial.

The following equation is used to update the bias weights of all the units in CA3 and DG at the end of each trial during recall.

$$dW_i = -(\epsilon * y_j) - (\alpha * W_i) \tag{4.1}$$

In equation 4.1, W_i is the bias weight for unit i, and dW_i is the amount that the weight is changed on each trial. ϵ determines the rate of weight change, and is set to 0.05. Change in bias weight scales with the unit's activity (y_j). α determines the rate of weight decay, and is set to 0.2 in these simulations.

k-Winners-take-all Inhibition (KWTA)

A k-Winners-take-all algorithm is used to control activation values in the network, which also causes the layers to have sparse distributed representations. This algorithm applies a uniform inhibition to all the units in the network. The inhibition level is set such that only k units will be active. Table 4.1 describes the percentage activity for the units in the various layers of the network.

A modification from the standard leabra++ algorithm (O'Reilly & Munakata, 2000) was applied in these simulations. There are two types of inhibition in these simulations. There is a bias term (described above) representing inhibition internal to each unit, and there is global inhibition that scales with local network activity. This global inhibition acts as a set-point for activity level in a given area. The strength of the inhibition scales with the amount of excitatory input to the units, causing the same number (k) of units to be active regardless of the overall excitation of the area.

As mentioned, each unit receives two types of inhibition, bias and KWTA. The bias inhibition is specific to each unit, as described above, while global inhibition scales with the excitation of the entire area. The standard leabra++ algorithm applies the global inhibition after the individual biases are applied. Thus, if the bias to a given unit has been made more negative, but the unit is still one of the k most excited units, the KWTA algorithm will cause this unit to still be one of the most active. For the current simulations, the order of application of these two types of inhibition was reversed. Thus, if the bias term for a particular unit was made more negative, that unit immediately began to decrease its activation.

Weight modification

Norman and O'Reilly (2003) contains details of the weight modification equations used in the current simulations. The error-driven component of weight change was turned off during all simulations presented here. However, the weights connecting layers EC and CA1 are pre-trained using the error-driven algorithm, in order to produce an invertible mapping of EC patterns to CA1 patterns and back (McClelland & Goddard, 1996; O'Reilly & Rudy, 2001).

The hippocampal component of the model implements Hebbian weight change, as described in several previous accounts of the model (O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003; O'Reilly & Munakata, 2000). Equation 4.2 describes the Hebbian weight change algorithm, applied at the end of the final plus phase. The term dw_{ij} represents the change in the weight connecting units i and j. The activation of the presynaptic unit is represented by x_i^+, and the activation of the postsynaptic unit is represented by y_i^+. The current weight between the units is represented by w_{ij}.

Area	Units	Activity (%)
posterior	120	10.0
EC	240	10.0
DG	1600	1.0
CA3	480	4.0
CA1	640	10.0
PFC	120	10.0
BG	120	10.0

Table 4.1: Number of units and average activity levels for the layers in the network. EC = entorhinal cortex, DG = dentate gyrus, CA3 and CA1 = hippocampal subregions, PFC = prefrontal cortex, BG = basal ganglia.

This term prevents the weights from growing without bound.

$$dw_{ij} = y_j^+ \left(x_i^+ - w_{ij} \right) \tag{4.2}$$

Other network and simulation parameters

Table 4.1 describes the number of units in each layer of the network and well as the activity level enforced by the KWTA algorithm.

Results are obtained by Monte Carlo simulation of the model. Each set of results consists of running the model on 300 free recall lists, except where noted below. After every run, the model was reinitialized. These 300 runs are grouped into sets of 10 (to simulate running 30 subjects with 10 lists apiece), in order to calculate error bars and statistics comparable to a behavioral experiment. Unless otherwise noted, standard error is calculated across this set of 30 simulated subjects. The results reported in section 4.3.3 are based on sets of 280 simulations.

4.2.3 Posterior cortex component details

The posterior layer of the network consists of 120 units, arranged in groups of 10 (for ease of creating item patterns: section 4.2.6). The posterior layer is connected to the entorhinal input layer, as well as the PFC and basal ganglia (BG) layers. The connection between the posterior layer and the entorhinal input layer is one-to-one; that is, there is a set of 120 units in the entorhinal input layer that only receive a connection from their corresponding posterior unit. The connection from

the posterior layer to the PFC and BG layers is 1:all. That is, each posterior unit connects to all of the PFC and BG units, although the strength of the connections are randomized upon network initialization. The posterior layer receives connections from the entorhinal output layer, in a similar 1:1 mapping.

Activity patterns are externally presented to the posterior layer during the training period / study period (section 4.2.6). During the recall period, I simulate pattern completion processes in the posterior layer as follows. The pattern of activity reinstated in the posterior layer by the connections from the entorhinal layer is compared to all of the studied patterns using a cosine comparison. If the similarity of the retrieved pattern to the any of the studied patterns exceeds 0.5, then the original pattern with the greatest cosine similarity to the retrieved pattern is reactivated in the posterior layer at the start of the subsequent trial.

4.2.4 Hippocampal component details

Details of the hippocampal model appear in Norman and O'Reilly (2003; O'Reilly & Rudy, 2001), including details of the connectivity of the sub-regions. In this section I describe the interactions of this component of the model with the posterior and prefrontal components.

The entorhinal input layer has two groups of units. One receives 1:1 connections from the posterior layer (units 1-120), the other receives 1:1 connections from the PFC layer (units 121-240). The entorhinal output area has the reverse pattern of connectivity.

During encoding it is important for the system to assign distinct patterns to each memory trace. However, during retrieval, it is important to find stored traces that are similar to the pattern currently active in the Entorhinal input area; in this case it is important to reduce the amount of pattern separation. By conjecture, I allowed the relative influence of the Dentate Gyrus layer to increase during encoding, and decrease during retrieval, satisfying these opposing constraints.

The Hebbian weight change equation, described in section 4.2.2, is applied to the following sets of weights: EC to DG, EC to CA3, CA3 recurrents, and CA3 to CA1, during the study period only.

4.2.5 Prefrontal component details

The prefrontal layer contains 120 units, arranged in groups of 10. Each set of 10 units is referred to as a stripe. A stripe can be "locked" or "unlocked". Every unit in a stripe has a hysteresis current that, when activated, acts as a tonic excitatory influence to that unit (Frank et al., 2001). Thus, when a stripe is locked, each unit's hysteresis current is set to the unit's activity value at the time of locking, causing the pattern of activity in that stripe to be maintained until the stripe is unlocked. When a stripe is unlocked, all of the hysteresis currents are set to zero.

The basal ganglia (BG) layer controls the gating in the PFC layer. Each stripe has a set of 10 units in BG that controls whether it is locked or not. Each set of 10 BG units contains two pools of 5 units, a "go" pool, and a "no" pool. During each trial, there is a competition between these units. Whichever pool has the maximally active unit wins the competition. The winning pool determines what happens to the corresponding PFC stripe. If the go pool wins, the lock-state of the PFC stripe is toggled. If the no pool wins, nothing happens.

Given a successful retrieval of an item by the network, the system updates PFC with retrieved context from the entorhinal output layer. Each stripe in PFC has a 50% chance of being updated by the corresponding units in the entorhinal output layer. It is this component of PFC gating that is damaged in section 4.3.4, by decreasing the probability of PFC update given a correct recall (see section 4.2.6).

4.2.6 Simulation of the free recall paradigm

Creating the patterns

A set of algorithms were created to automatically generate the study patterns comprising a given free recall list. The posterior area of the model is comprised of 12 slots; each slot has 10 units, which can be on or off for a given input pattern. For a given input pattern, each slot is assigned a single "on" unit to create a pattern. The patterns are not constrained to be orthogonal, the "on" unit for each slot for a given pattern was chosen at random.

Training period

For the simulations presented in sections 4.3.2 and 4.3.4, the training period consisted of 20 items presented serially to the network, on successive trials. Each item pattern was only presented once. For the simulations presented in section 4.3.3, lists of 12 items were used, to match the list length of the behavioral experiment reported in chapter 3.

Recall period

During the recall period of the simulation, there were forty recall trials. During each trial, the hippocampal component of the model was probed with a pattern of prefrontal activity. A successful recall occurred when reinstated activity in the item component of the EC output area matched a studied pattern (section 4.2.3). Given a successful recall, two things happened. First, some subset of the PFC stripes were updated with the retrieved pattern of activity in the context component of the EC output area (section 4.2.5). Second, the item representation corresponding to the retrieved item was reinstated in the posterior component on the start of the next trial (section 4.2.3).

Scoring recall performance in the model

During the recall period, a set of patterns are retrieved by the network. Only a subset of these patterns were considered valid recalls. For example, retrieved patterns that failed to exceed a similarity threshold (section 4.2.3) to a studied pattern were excluded. Furthermore, patterns that had already been recalled during that recall period were excluded. After these exclusions were applied, the first recall in the list was used for the probability of first recall (PFR) curves. This list of recalls was also used to calculate transition probabilities for the conditional response probability (CRP) curves. See Kahana (1996) for details of this calculation. In short, the CRP for a transition of a given lag is the number of transitions made of a given lag divided by the total number of possible times a transition of that lag could have been made (for example, given a recall of the second-to-last item on the study list, a transition with lag of +1 is possible, but +2 is not). Transitions to serial positions of already-recalled items were not considered possible.

4.2.7 Simulation of encoding task performance

Section 4.3.3 describes a set of simulations designed to mirror the methods followed in the behavioral experiment described in chapter 3. In order to simulate task performance, two orthogonal task representations were created. Each of the representations consisted of a pattern of activity across 5 PFC stripes, in which one unit was activated per stripe. The two task representations were orthogonal.

As described in section 4.4.1, the current version of the model lacks a mechanism for ending the recall attempt. Whereas human subjects tend to give up after some number of attempts, the current model continues to recall throughout the recall period. Preliminary simulation work suggested that a shorter recall period provided a closer fit to behavioral data. For this set of simulations the recall period was shortened significantly, to 10 recall trials. Future work will investigate alternate mechanisms for recall termination.

4.2.8 Simulation of brain damage

Three types of damage are applied to the model in section 4.3.4, in order to fit the pattern of behavioral deficits seen in a population of elderly subjects in free recall (Kahana et al., 2002). The first two types of damage simply removed a set of units from certain components of the model before training began. The final type of damage altered the gating mechanism described in sections 1.3 and 4.2.5.

Broad lesions of the hippocampal and prefrontal components

To simulate lesions to brain regions, a proportion of units in subcomponents of the model were removed before the simulation began. To simulate hippocampal lesions, an algorithm randomly selected 70% of the units in the DG, CA3, and CA1 layers of the network and removed these units from the simulation. To simulate PFC damage, this algorithm removed 70% of the PFC units (none of the BG units were removed).

Simulating gating failures

To simulate damage to the prefrontal gating system, I introduced gating failures into the algorithm that allowed information from the EC output area to be gated into PFC. As described above (section 4.2.6), this updating algorithm is run when a correct recall was made. In the damage condition, this updating process could randomly fail, causing no information to be transferred from EC to PFC. The case explored in the text (and presented in Figure 4.6) had a 50% chance of updating failure.

Fitting CRP curves

In order to determine whether the CRP curves reported in the gating damage simulation (section 4.3.4) significantly differed from baseline, I followed a technique described in Kahana et al. (2002). For this procedure, a power function (Eq. 4.3) was used to fit the forward legs of each set of CRP curves. In this equation, $CRP(lag)$ corresponds to the height of the CRP curve at a given lag (i.e. the probability of a given transition).

$$CRP(lag) = a * lag^b \tag{4.3}$$

As mentioned above (section 4.2.2), simulations were grouped in sets of 10. Each of these sets produced a CRP curve. The forward leg of each CRP curve was fit with the power function, producing a set of exponential fits (the b term).

4.3 Results

4.3.1 Overview

In section 4.3.2 I describe simulations of the basic free recall paradigm. A serial position curve of model output shows that recent items are more likely to be recalled than early list items (Figure 4.1). However, the model does not fit the primacy and mid-list asymptote seen in the serial position curve of human subjects. The model does fit the basic shapes of the PFR and CRP curves (Figure 4.2).

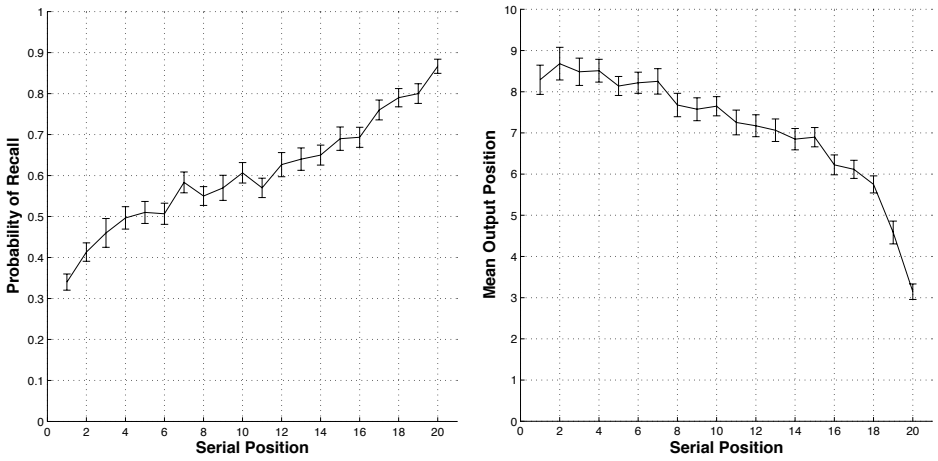
In section 4.3.3 I describe simulations of task switch during the study list. The model shows enhanced recall of post-switch items but also shows a decreased probability of recall for pre-switch items (Figure 4.3). The model shows a reduced probability of making an output transition that crosses the task switch boundary, relative to the control condition (Table 4.2).

In section 4.3.4 I describe simulations attempting to fit elderly performance in free recall, using the simulations described in section 4.3.2 as baseline (younger subjects). Damage to the hippocampal component of the model (lesion of 70% of the units) causes flattening of both the PFR and CRP curves (Figure 4.4). Damage to the prefrontal component of the model (lesion of 70% of the units) also causes flattening of both the PFR and CRP curves (Figure 4.5). Finally, damage to the prefrontal gating component of the model caused the desired pattern, a spared PFR curve and a flattened CRP curve (Figure 4.6).

4.3.2 Applying the model to free recall

As mentioned in section 4.2.6, a set of 300 simulations were run to generate the results in this section. The average percentage of studied items recalled was 61% with a standard error of 0.65%. Figure 4.1 shows the performance of the model on two canonical measures of free recall performance, the serial position curve and the mean output position curve. The serial position curve depicts the probability that an item from a given position in the study list will be recalled. The mean output position curve gives a sense of the order of recall by plotting the mean output position for every serial position in the study list. These two curves give a good first-pass description of the behavior of the model in retrieving patterns. Figure 3.2(a) shows a comparable serial position curve drawn from the control condition of the behavioral experiment described in chapter 3. As is evident from comparing these figures, the simulated serial position curve does not capture the primacy effect or the mid-list asymptote so often seen in the behavioral literature (Parkin, 1993). Possible reasons for this will be elaborated in section 4.4.1.

From the serial position curve (Fig. 4.1(a)), it is clear that overall, the model is more likely to recall items from the end of the study list. Furthermore, the mean output position curve (Fig. 4.1(b))



(a) The serial position curve

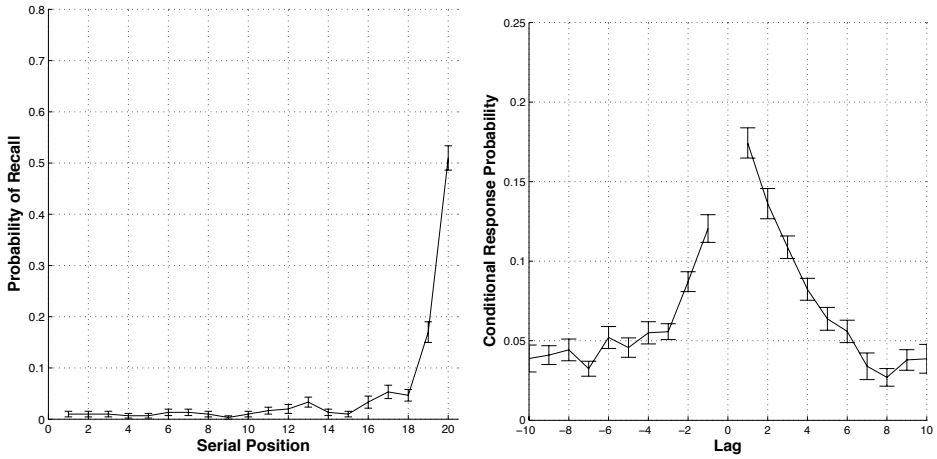
(b) The mean output position curve

Figure 4.1: Probability of recall by serial position during study. Mean output position by serial position during study. See text for details.

shows that the model, like humans, tends to recall end-list items earlier in the recall process than items from early serial positions. These effects are due to the slow drift of the PFC state over the course of the study list, coupled with the fact that the model initiates recall with a PFC state similar to that present while the end-list items were studied.

Figure 4.2 shows two other important measures for characterizing free recall performance, the probability of first recall (PFR) curve and the conditional response probability (CRP) curve. These curves characterize the output transitions that a subject makes during free recall (see chapter 1). Figure 4.2(a) characterizes how the model enters the list; it tends to initiate recall with an end-list item, most likely the final item. In this regard the model matches behavioral results (Kahana, 1996). For comparison with behavioral data, refer to Figure 1.1 (the "Immediate" recall condition).

The CRP curve (Fig. 4.2(b)) characterizes the transitions that a memory system makes from item to item during recall. Items studied nearby in time tend to be recalled on successive retrieval attempts. This is because on successive recalls, the memory system is probed with similar PFC states, which tends to draw out items studied nearby in time. For comparison with behavioral data,



(a) The probability of first recall curve

(b) The conditional response probability curve

Figure 4.2: Probability of first recall by serial position. Conditional response probability.

refer to Figure 1.2 (the left-hand side, "Young" subjects).

4.3.3 Task simulations with the model

As described in section 4.2.7, task representations were simulated as static patterns in the PFC component of the model. In this section I describe the effects of changing this task representation halfway through the study list on model behavior during the recall period. Figure 4.3 shows recall performance broken down by list half. The green line shows recall performance in the control condition. In the behavioral results (Figure 3.1(c)), subjects show first-half recall performance that is comparable to second-half recall performance. Possible reasons for this are presented in section 4.4. The red line in Figure 4.3 shows model performance in the halfway switch condition. The model shows a recall advantage for second-half items, and a disadvantage for first-half items. Thus, there is a cost, as well as a benefit for the halfway switch condition. This is consistent with certain context change paradigms (such as directed forgetting or environmental context change, see section 1.2.1), but is not consistent with the behavioral results in Figure 3.1(c), in which there was no effect of encoding task switch on first-half recall. Possible reasons for this are discussed in

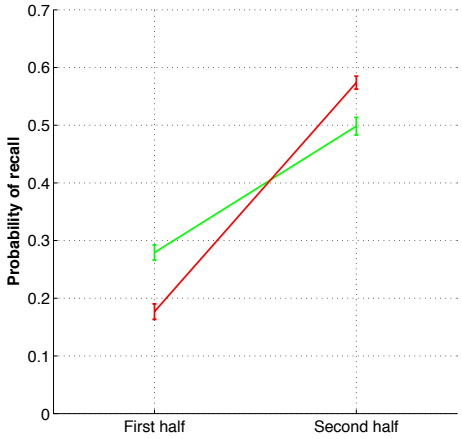


Figure 4.3: Simulation of task performance using the model. Probability of recall broken down by both list type and list half. The green line corresponds to the control condition, and the red line corresponds to the switch condition. For comparison with behavioral data, see Figure 3.1(c).

	Cross-half transition probability	Standard error
Control	0.4241	0.0204
Switch	0.1695	0.0133

Table 4.2: The probability of a cross-half output transition broken down by list type. For comparison with behavioral data, see Table 3.1. The model is much less likely to make a cross-half transition in the switch condition. A two-sample t-test (assuming unknown and unequal variances) was used to compare the sets of transition probabilities ($p<0.0001$).

section 4.4.2.

Table 4.2 shows that there is a significant difference in the probability of a cross-half transition for control versus halfway switch. This matches the behavioral result shown in Table 3.1. In section 4.4.2, I elaborate upon the correspondence between the simulation results and the behavioral results, in terms of context-based theories of memory retrieval.

4.3.4 Comparing the model to healthy aging data

As mentioned above, one advantage of the present model is that it ascribes function to various brain regions, making it possible to investigate the extent to which damage to an area of the model produces behavior that matches the behavior of a special population. The neuropsychological literature is a large one, but unfortunately, most investigations of special populations in free recall do

not report the detailed output transition measures that our model seems to fit well.

The current investigation is restricted to a study of free recall (Kahana et al., 2002; Howard et al., submitted) in which the performance of young and older subjects were compared on a variety of measures, including transition probabilities. I applied several types of damage to the model, in order to determine whether I could capture the pattern of deficits seen in a healthy aged population.

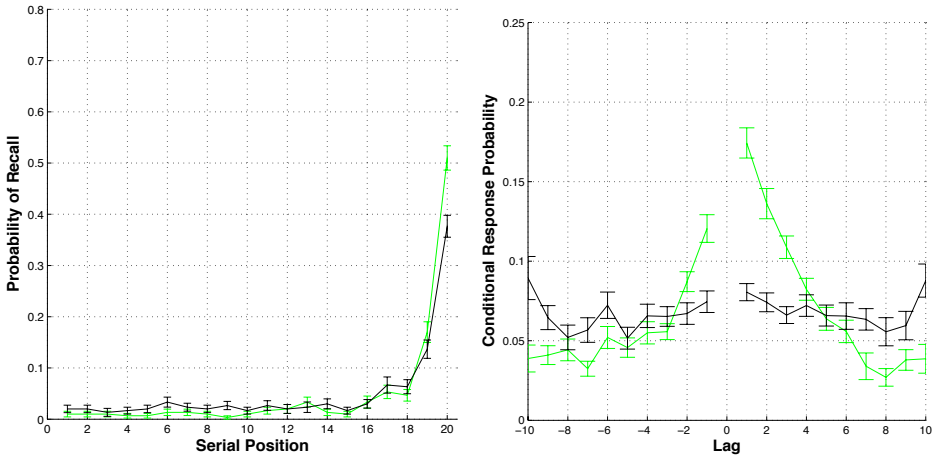
The pattern of deficits is best described as such: Elderly patients showed a reduced probability of recall at all serial positions and total recall was significantly reduced. Furthermore, the output transitions that were made during recall showed an altered profile; the CRP curve was significantly flattened in both the forward and backward directions. The most constraining point is that the PFR curve appeared unaltered; elderly subjects showed the same probability profile as the young with regards to initiation of recall.

The first investigations looked at simple types of damage meant to correspond to cell death. In these investigations I simply removed a large number of units from various components of the model and examined the pattern of results. Follow up analyses, to be described below, investigated more subtle types of damage to the system.

Hippocampal damage

In this first study, I investigated the effect of broad damage to the hippocampal subsystem on the performance of the model. I removed 70% of the units from the hippocampal subsystem (section 4.2.8). The model is remarkably robust to this type of damage (in that it still works at all). However, the hippocampal lesion substantially affected the pattern of results obtained in the free recall paradigm.

The conditional response probability graph was completely flattened (Figure 4.4(b), black line) relative to the baseline condition (green line). The hippocampal component of the model was retrieving a degraded version of context; as such, the transitions made by the model did not follow the pattern of lag-recency. Furthermore, the probability of first recall measure shows a significant flattening for the last serial position (Figure 4.4(a), black line). This does not fit the pattern seen in



(a) The probability of first recall curve

(b) The conditional response probability curve

Figure 4.4: These figures depict model performance after damage to the hippocampal component. 70% of the hippocampal units were removed. The black line shows damaged model performance. The green line shows intact model performance, for reference.

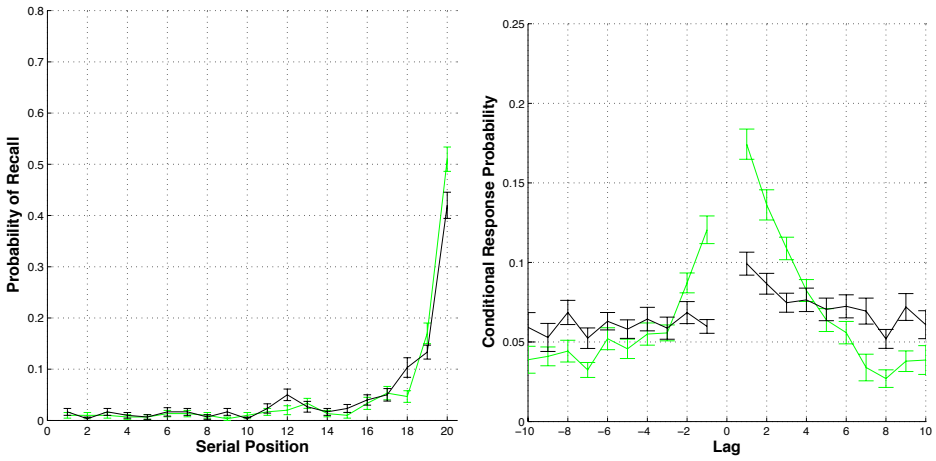
the elderly subjects.

Prefrontal damage

A similar lesion was applied to the prefrontal component of the model; 70% of the units were removed. Again, the conditional response probability measure showed considerable flattening (Figure 4.5(b), black line) relative to the baseline condition (green line). So did the probability of first recall measure (Figure 4.5(a), black line), which again, does not fit the measured pattern seen in the elderly subjects.

Exploring gating failures

These brute types of lesions to the model do not allow it to capture the profile of the elderly subjects. However, there is another class of damage to consider, damage to the gating system of the model. Here, I drew inspiration from the above mentioned study of the elderly in task performance (Braver et al., 2001, described in section 1.2.3); this study suggests that the elderly cognitive system



(a) The probability of first recall curve

(b) The conditional response probability curve

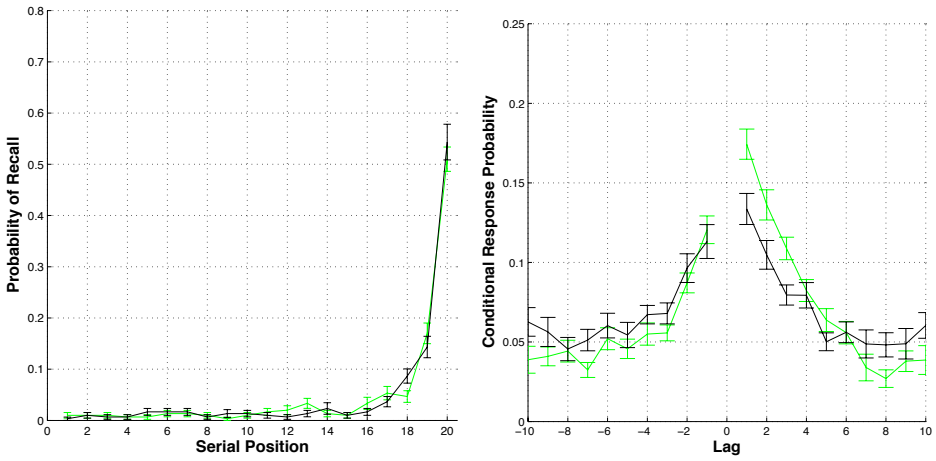
Figure 4.5: These figures depict model performance after damage to the prefrontal component. 70% of the prefrontal units were removed. The black line shows damaged model performance. The green line shows intact model performance, for reference.

has difficulty gating new information into prefrontal cortex.

There are two important periods in the experiment when the gating system causes the PFC component of the model to update inner mental context. The first is during the study period, while patterns are being presented for encoding. Here, each new pattern drives the model to open a subset of gates, which allows input pattern activity to influence the state of PFC. The second period in which the gating system becomes important is during recall, when patterns of activity retrieved by the memory system are gated into PFC, which alters the memory cue and drives further recall. In the first case, the gating is driven by an event occurring in the external environment. In the second case, the system is relying on an internal recall event to drive gating. I made an assumption that in the second case, the signal driving gating would be much weaker, and constructed a simulation in which the gating system functions normally when driven by items in the external environment, but fails to update 50% of the time during the recall process. The motivation for this will be discussed further in section 4.4.3.

In this damaged gating system simulation, the conditional response probability measure was altered (Figure 4.6(b), black line). The graph shows significant flattening, as described below. However, there is no effect on the probability of first recall measure (Figure 4.6(a), black line). This simulation fits the pattern of results seen in the elderly subjects.

In order to quantify the significance of the CRP flattening in the damaged gating system simulation, a power function was used to fit the forward legs of each set of CRP curves (section 4.2.8). Here, the two sets of curves that were fit were drawn from the set of baseline simulations, and from the set of damaged-gating simulations (the green curve and the black curve in Figure 4.6(b), respectively). Equation 4.3 was used to fit the forward-leg curves. The mean best-fit parameter for the exponent (b) was -0.6483 for the baseline simulations, and -0.4578 for the damaged-gating simulations. An unpaired two-tailed t-test (assuming unequal variance) was performed on the two sets of fitted exponents. The impairment in the damaged-gating simulations was statistically significant, $t(57) = -3.29$, $p<0.005$.



(a) The probability of first recall curve

(b) The conditional response probability curve

Figure 4.6: These figures depict model performance after damage to the gating system. The context system was only allowed to update 50% of the time during the recall period. The black line shows damaged model performance. The green line shows intact model performance, for reference.

4.4 Discussion

4.4.1 Comparing the model to behavioral data

This tripartite model of the interactions of a semantic system, memory system and context system captures the rich behavior of subjects in a free recall paradigm. Major features of human behavior are captured, but there are some features of the system's performance that do not match the human profile.

For example, the model does not yet capture the primacy effect (seen as an enhanced probability of recall of the initial list items). I believe there are two reasons for this. First, a major component of the primacy effect can be explained by the tendency of subjects to rehearse early list items to a greater degree than mid-list items. The current version of the model has no mechanism for rehearsal; as such, all items get equivalent encoding. However, rehearsal can not be the only explanation of the primacy effect. In the continuous distraction version of the free recall paradigm, subjects still show a pronounced effect of primacy. Presumably, this primacy effect is somehow produced by the interaction of the context and memory systems. It has been shown behaviorally that subjects, in the middle of the recall period, tend to make a transition to the first item of a list (Laming, 1999). Laming (1999) explains this by positing the existence of a von Restorff stimulus at the beginning of the list, corresponding to the start signal given to the subject.[1] This von Restorff stimulus is supposedly more strongly encoded than the other stimuli, leading to an enhanced probability of recall of the first item, which is encoded closest in time to the start signal.[2] Once the first item is recalled, presumably the subject is biased to make forward transitions to local items, extending the primacy benefit to the first few items. The current version of the model does not include either a rehearsal mechanism or a von Restorff-like enhancement mechanism, which may explain the model's inability to capture the primacy effect.

[1] A von Restorff stimulus is an item or event that stands out from the surrounding events. A classic example of this would be a digit in a list of letters. While there is no requirement to recall the start signal of a given list, Laming (1999) suggests that it is covertly recalled, and then subjects transition to an early list item.

[2] One simple way to implement this sort of enhancement would be to transiently increase the learning rate in the hippocampal component of the model for the first item.

The current model does a good job fitting the basic shape of the CRP curve seen in human behavior. See Figures 4.2 and 1.2 for model and human performance, respectively. The mechanism by which the current model shows a forward bias in output transitions (Polyn, Norman, & Cohen, 2003) is similar to that proposed by Howard and Kahana (2002a). As described in section 1.2.2, there are two sources of retrieved context in the TCM model, pre-experimental and experimental. In the current model, pre-experimental context is represented by the direct connection between the posterior component and the PFC component. This is the path by which item-specific features perturb the PFC representation during study, as well as the path by which these same item-specific features are reinstated in PFC during recall of a given item representation (see section 1.3 for more detail on this process). Just as in the TCM model, pre-experimental context is responsible for the forward bias in the CRP curve. Some preliminary simulation work (results not shown here) bears this out. If the connection from the posterior component to the PFC component of the model is turned off during recall, the CRP curve becomes symmetric. In this case the PFC representation is only updated by the connection from the entorhinal component of the model (acting as experimental context).

Finally, in comparing the current set of simulations to those generated by models like SAM, which have fleshed out the algorithmic aspects of memory search to a greater degree, it becomes clear that there are some shortcomings of the model in this domain. In the SAM model, there is a mechanism by which subjects terminate the recall process after some number of retrieval attempts that fail to retrieve an not-previously-recalled item. The current model lacks such a mechanism. For example, even with the unsticking mechanism, the current model tends to continue to recall items across the entire recall period, often returning to items that were recalled earlier. Thus, the current version of the model makes no predictions about the point at which subjects will terminate the recall process. This may explain why the serial position curves generated by the model do not match empirical results. Preliminary simulations, in which the length of the recall period is greatly shortened, produce serial position curves that asymptote at a low level of recall for early and mid-list items. These serial position curves still do not show the primacy effect, but possible reasons for that

are discussed above.

4.4.2 Encoding task performance in the model

In section 4.3.2, I described results of simulations in which the identity of the encoding task was represented in the PFC component of the model. This allowed me to simulate the encoding task switch paradigm described in chapter 3, by changing this representation halfway through the study list (see section 4.2.7 for details).

As described, the model broadly matched the behavior of the human subjects. It captured the increased performance on the second half of the list in the halfway switch condition, as well as the reduced probability of a cross-half output transition. However, the model performed worse on first half recall in the halfway switch condition, relative to the control condition. In contrast, the human subjects performed equivalently on first half recall across the switch and control conditions. One possible explanation for this lack of a cost of encoding task switch on recall performance is that the human subjects were able to reinstate the contextual representation corresponding to the first half encoding task, driving relatively normal recall for these items. This suggests a modification to the model in which task context can be reinstated during recall (equivalent to the subject thinking broadly about the encoding task in order to probe memory for stored traces).

While the current version of the model can reinstate task context through the retrieval operations of the hippocampal component, the representation of task is not treated any differently than the rest of the PFC representation. One might imagine that there are other mechanisms for learning the general character of operations performed during the study period. For example, Becker and Lim (2003) (reviewed in section 1.2.2) propose a learning mechanism in the PFC that allows their model to rapidly acquire an internal representation of the categories presented during the list. Their model is then able to reinstate these category representations during the recall period, driving retrieval of the studied items from each category. A mechanism like this might allow the model to more reliably reinstate the pre-switch task context, driving reliable recall of the first half items. This would also allow the model to make connections to the literature on learning task representations (Cohen et al.,

1990; Posner & Snyder, 1975).

As shown in Table 4.2, the model was able to capture the 'memory isolation' effect reported in Table 3.1, in which subjects were less likely to make a recall transition across the list half boundary when the two halves of the list were encoded with different tasks. In the halfway switch condition, each set of studied items had a different task representation associated with it. At the beginning of recall, the PFC representation still had the second encoding task representation active, driving recall of items from the second half of the list. Over the course of recall, this task representation would degrade, increasing the probability that the model would retrieve an item from the first half of the list. Further simulations will be necessary to explore the dynamics of these task representations during the recall period.

4.4.3 Capturing the pattern of deficits in the elderly

While the current results are suggestive regarding the damage to the cognitive system in the elderly, this problem must be approached carefully. In the simulations described in section 4.3.4, the gating system reacts differently to an externally-driven and an internally-driven gating signal. That is, upon presentation of an item during the study period, the PFC component updates normally. However, given recall of a hippocampal memory trace, the PFC component has a reduced probability of allowing retrieved representations to be reinstated. Thus, in these two cases the source of the gating signal is different, arising from posterior cortex during study, and hippocampus during recall. It is possible to imagine that these are dissociable components of the gating system. However, convergent validation from other domains would make the story more plausible.

It is interesting that the TCM model and the current model diverge in their explanations of the pattern of deficits in the elderly in free recall. In each model, the context vector is updated by two sources during retrieval (pre-experimental and experimental). As described in section 1.2.3, the TCM model fits the data by reducing the influence of experimental context at recall. Given recall of the item from serial position N, experimental context provides a symmetric cue for items in both directions (N-1 and N+1). By reducing the effect of experimental context in the TCM

model at recall, the CRP curve is flattened, but retains the forward asymmetry due to the retrieval of pre-experimental context. By the TCM story, experimental context is supplied by the hippocampal component, so reducing the influence of experimental context implies degraded hippocampal output.

In the current model, there are two components responsible for the reinstatement of experimental context. First, the hippocampal component retrieves a memory trace that contains both item and (experimental) context representations. Then, the gating system allows the retrieved experimental context to be reinstated in the PFC component. When the hippocampal component of the current model is damaged, it hurts the ability of the model to reinstate context, but it also hurts the ability of the model to reinstate item representations, including the very first item recalled (which flattens the PFR curve; Fig. 4.4(a)). However, when the gating system is damaged, the hippocampus retrieves a faithful context and item representation, but the system is unable to transfer that context representation to the PFC component. The PFR curve is unaffected (Fig. 4.6(a)), because the first recall is cued by the end-of-list PFC representation. In the damaged gating simulations, subsequent recalls are affected, as the system is unable to properly update context using representations coming from the hippocampal component.

The current explanation and the TCM explanation of the pattern of deficits in elderly subjects have much in common. In each case an analogous component of context has a reduced influence during recall. However, structural differences between the models change the interpretation of the results. In the TCM model, experimental context representations are retrieved by the hippocampal component, and then the updated context representation directly retrieves an item. Thus, in the TCM model, retrieval of item representations is not mediated directly by the hippocampal component, so damage to this component does not harm retrieval of the first item. However, in the current model, the hippocampal component is responsible for retrieval of both item and context information, so damage to this component harms retrieval of the first item. In the current model, the reinstatement of experimental context requires two intact systems, hippocampus to retrieve it, and the gating

system to maintain it. Thus, by damaging the gating system, I was able to capture the observed pattern of deficits for older subjects in free recall.

Chapter 5

General Discussion

5.1 Introduction

Free recall is not an easy paradigm to explain. It involves the interaction of all the major cognitive systems, and researchers and theoreticians have been generating data and theory for at least a century. In wading through these behavioral findings and purported mechanisms, the value of a formal model in organizing thought and extracting principles of function of the broader cognitive system has become apparent to me.

In this dissertation, I presented three investigations of the human memory system in free recall. In chapter 2, I described an imaging experiment in which patterns of brain activity were related to memory targeting processes. In chapter 3, I described a behavioral experiment in which I examined the relationship of task representation and memory accessibility. In chapter 4, I described simulation studies of a computational model that uses context representations to probe a memory system for stored patterns. In this chapter I describe some of the ways that these projects intersect and future directions for the overall endeavor.

5.2 Points of contact and future directions

5.2.1 Imaging memory targeting

There are two issues I would like to discuss regarding the future direction of this imaging work. The first regards a debate in the literature about the functional nature of activity patterns recorded in the domain of categorical perception. To set up the debate, consider the fusiform face area, an area of fusiform cortex that tends to activate strongly and reliably when a subject views, remembers or otherwise processes a face (Kanwisher, McDermott, & Chun, 1997; O'Craven & Kanwisher, 2000). How are we to interpret this activation, with regards to the function of the broader cognitive system? One proposal is that this region of cortex represents a face module, centrally dedicated to the processing of faces (Kanwisher et al., 1997). Another proposal is that this fusiform activation is simply the peak of a broadly distributed and highly textured pattern of activation, all of which represents processing of the currently viewed face (Haxby et al., 2001; Hanson et al., 2004). Thus, the issue at the center of the debate is whether the representation of "face" (or any other category) is distributed across the ventral temporal lobe, or is localized to a small module within it. The current data can be inspected to see whether the signal in canonical category-sensitive areas correlates with behavior as well as the distributed pattern. However, to truly resolve this debate, it will be necessary to show that category-related fluctuations in the broader activity pattern do not just correlate with behavior, but are causal to behavior (Cohen & Tong, 2001).

A follow-up analysis will develop masks based on the anatomical areas which a priori are thought to contain these category-selective modules (for faces, locations and objects). By training the classifier on the union of these masks, it should be possible to establish whether the classifier output better correlates with behavior when just given these core modules, or when it is given the full distributed pattern of brain activity. The results reported in section 2.3.8 already lean toward the distributed interpretation. The classifier trained on only temporal lobe data performed worse (in terms of correlation with verbal recalls) than the classifier trained on whole-brain data. Thus, areas outside the temporal lobe provided valuable information which allowed the classifier to discriminate between

the three categories. Given these results, I expect that the classifier will do worse still when trained on a subspace of the temporal lobe containing the face, location and object areas.

The second issue regards the future of this line of experiments. In the current experiment I established that one can estimate the strength of a pattern of activity related to memory targeting. The follow up experiment described here will attempt to tie the fluctuations of these patterns of brain activity to theories of memory arising from the cognitive psychology literature.

Baddeley (1990) presents a context-based account of proactive interference in the Brown-Peterson task (see section 1.2.1 for more details). In this task, subjects study a series of items and are asked to recall them after a period of distractor task performance. When multiple lists contain items all drawn from the same category, performance decreases over lists. In other words, a given cue facilitates retrieval of associated memories, but when too many memories are associated with a given cue, they begin to interfere with one another. However, when the study item category is changed (for example from types of fruits to names of professions), recall performance increases (Wickens, 1972). This phenomenon is known as release from proactive interference. Interestingly, more release is seen when the new category is quite different from the previous category (as above) than when the new category is relatively similar to the previous category (such as a change from types of fruits to types of vegetables).

I believe that pattern classification techniques of fMRI data can be applied in this domain. If categories of stimuli are chosen that evoke distinct patterns of brain activity detectable by these techniques, it should be possible to correlate the degree of release from proactive interference with the difference in brain patterns evoked by the two stimulus categories. In other words, a shift in brain pattern between lists implies that a new context representation is being used by the system. This new context representation should not be associated with memories encoded before the category shift, making it a better cue for memories encoded after the category shift. Interestingly, by this account, it is not the fidelity of the context representation that will determine recall performance, but the degree to which the current context representation differs from the previous context representation.

A number of metrics would be reasonable to use to quantify the difference between two stimulus category patterns, such as a cosine comparison or a correlation metric. If the degree of release from proactive interference correlates with the degree of shift in brain pattern, this would provide convergent evidence that these patterns are indeed being used to target memories. Furthermore, the current model could be applied to the Brown-Peterson domain, facilitating a quantitative comparison of brain data, behavioral data, and simulation results.

5.2.2 Behavioral investigations

As I mentioned in chapter 3 the bulk of studies that have investigated task effects in the free recall literature did so by manipulating the distractor task, the task performed between study of the items (Thapar & Greene, 1993; Neath, 1993). In this section I describe ongoing and future work that is designed to unify theories of encoding task and distractor task effects in free recall.

Carl Gold (a senior at Princeton University working with Professor Ken Norman) and I designed an experiment that mirrors the one described in chapter 3. In this experiment, subjects were run in a continuous distractor free recall paradigm. There were two conditions, one in which distractor task switches halfway through the list, and another in which the same distractor task is performed throughout the list. The encoding task remained constant, and was the same size judgment described in section 3.2.4. The distractor tasks were an asterisk counting task and a gender naming task, both drawn from Thapar and Greene's (1993) study of distractor task switching.

There was no effect of distractor task switch on the serial position curves of the experiment. However, there was a small but significant decrease in the probability of a subject making a transition across the list half boundary. It is interesting that the effect of distractor task switch was so subtle, given the larger effects reported by Thapar and Greene (1993). One difference between the two paradigms is that in the one run by Gold, the study items were all encoded with the same task. In the Thapar and Greene (1993) study, there was no unifying study task. Thus, it is possible that by weakening the depth of the encoding task, the distractor task may have more influence on memory accessibility. This interplay between task context and memory accessibility seems like a rich vein

for future work.

5.2.3 Future simulation work

There are a number of potential directions for development of the current model. The first step is to apply standard simplex parameter fitting algorithms to maximize the fit of model output to behavioral data. This procedure will reduce the amount of hand-tuning of the network, increasing confidence that the final parameter settings are not some sort of local minimum in the high-dimensional parameter space.

Part of the power of the current model is in its ability to fit data from multiple domains. The model has successfully been applied to the domains of recognition memory (Norman & O'Reilly, 2003), source memory (Polyn et al., 2002) and free recall (Polyn et al., 2003 and chapter 4). There are a number of related paradigms that deserve exploration. In section 1.2.1 I described a contextual account of results from the directed forgetting paradigm. The current model could be applied to data from this domain.

In the directed forgetting paradigm, presentation of a cue to forget a set of items, under certain circumstances, causes subjects to have poorer memory for those items (Geiselman et al., 1983). In section 1.2.1 I described the directed forgetting behavioral phenomena in terms of changes to an inner mental context representation. This explanation can now be elaborated in terms of the current model, as a guide for future simulation work.

As mentioned in section 1.3, the PFC component of the model contains a representation that is slowly updated by the presentation of new study items. This updating is carried out by a set of gates that are either locked, causing activity patterns to be maintained in a given PFC subregion, or unlocked, allowing the PFC subregion to update upon presentation of new information. When a forget cue is given, a large number of the PFC gates unlock. Second set items then greatly perturb the PFC representation, shifting it to a new configuration that does not match the first half representation. This causes successive memory traces to be stored in a new region of memory space. At recall, the end-of-list PFC state will be a good cue for the items from the second set, but because the

PFC representation was greatly shifted after the forget cue, the first set items will be more difficult to access.

However, if any first set items are directly presented at test, they will allow the system to retrieve the PFC representation that preceded the forget cue. This will eradicate the deleterious effects of the forget cue. Finally, if no second set items are presented after a forget cue, inner mental context will not be perturbed, and recall will proceed normally.

In recent years, a number of models of free recall have been proposed (Howard & Kahana, 2002a; Becker & Lim, 2003; Polyn et al., 2003; Davelaar et al., 2005). Each of these models uses a different set of mechanisms to fit behavioral data. From a theory-development standpoint, it is very desirable to have multiple models that can be applied to the same data. For example, in section 1.2.3 I describe a TCM account of a pattern of deficits seen in the elderly in the free recall paradigm. In section 4.4.3, I describe how the current model fits the same pattern of deficits. In comparing the two accounts, it is clear that there are relevant structural differences between the models. While the hippocampal component of the TCM model only retrieves context information, the hippocampal component of the current model retrieves item and context information concurrently. A potential path for future work will be to look for convergent evidence for one or the other account in other domains. For example, the study by (Braver et al., 2001) suggests that damage to the prefrontal gating system could be responsible for the pattern of deficits seen in the elderly on a continuous performance task (the AX-CPT). If the current model could be developed to perform on both a free recall task and a version of the AX-CPT, it would be possible to investigate whether the elderly impairments in both tasks spring from a common cause.

5.3 References

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, *22*, 425–490.

Baddeley, A. (1990). *Human memory: theory and practice*. Boston, MA: Allyn and Bacon.

Baddeley, A. (1998). *Human memory: theory and practice, revised edition*. Boston, MA: Allyn and Bacon.

Becker, S., & Lim, J. (2003). A computational model of prefrontal control in free recall: strategic memory use in the california verbal learning task. *Journal of Cognitive Neuroscience*, *15*, 821–832.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: essays in honour of Endel Tulving* (pp. 309–330). New Jersey: Lawrence Erlbaum Associates.

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.), *Current issues in cognitive processes: the Tulane Flowerree symposium on cognition* (Chap. 9). New Jersey: Lawrence Erlbaum Associates.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173–189.

Bower, G. (1967). A multicomponent theory of the memory trace. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation*, Vol. 1 (pp. 229–325). New York: Academic Press.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton, & E. Martin (Eds.), *Coding processes in human memory* (Chap. 5, pp. 85–121). New York: John Wiley and Sons.

Braver, T. S., Barch, D. M., Keys, B. A., Carter, C. S., Cohen, J. D., Kaye, J. A., Janowsky, J. S., Taylor, S. F., Yesavage, J. A., & Mumenthaler, M. S. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology General*, *130*, 746–763.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell, & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*, 12–21.

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T. A., & Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Human Brain Mapping*, *12*, 61–78.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704–717.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45–77.

Cohen, J. D., & Tong, F. (2001). The face of controversy. *Science*, *293*, 2405–2407.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, *19(2 Pt1)*, 261–70.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*, 171–178.

Cox, R. W., & Jesmanowicz, A. (1999). Real-time 3d image registration for functional MRI. *Magnetic Resonance in Medicine*, *42*, 1014–1018.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*, 3–42.

Dobbins, I. G., Foley, H., Wagner, A. D., & Schacter, A. D. (2002). Executive control during episodic retrieval: Multiple prefrontal processes subserve source memory. *Neuron*, *35*, 989–996.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification, second edition*. New York: John Wiley and Sons.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377.

Ferino, F., Thierry, A. M., & Glowinski, J. (1987). Anatomical and electrophysiological evidence for a direct projection from ammon's horn to the medial prefrontal cortex in the rat. *Experimental Brain Research*, *65*, 421–426.

Fernandez, A., & Glenberg, A. M. (1985). Changing environmental context does not reliably affect memory. *Memory and Cognition*, *13*, 333–345.

Fischler, I., Rundus, D., & Atkinson, R. C. (1970). Effects of overt rehearsal processes on free recall. *Psychonomic Science*, *19*, 249–250.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience*, *1*, 137–160.

Geiselman, R. E., Bjork, R. A., & Fishman, D. (1983). Disrupted retrieval in directed forgetting: a link with posthypnotic amnesia. *Journal of Experimental Psychology: General*, *112*, 58–72.

Gelfand, H., & Bjork, R. A. (1985). On the locus of retrieval inhibition in directed forgetting. *Poster presented at the Meeting of the Psychonomic Society, Boston, MA*.

Gershberg, F. B., & Shimamura, A. P. (1995). Impaired use of organizational strategies in free recall following frontal lobe damage. *Neuropsychologia*, *33*, 1305–1333.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., Fish, J. H., & Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *6*, 355–369.

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 3–15.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology*, *66*, 325–331.

Goldman-Rakic, P. S., Selemon, L. D., & Schwartz, M. L. (1984). Dual pathways connecting the dorsolateral prefrontal cortex with the hippocampal formation and parahippocampal cortex in the rhesus monkey. *Neuroscience*, *12*, 719–743.

Greene, R. L. (1992). *Human memory: paradigms and paradoxes*. New Jersey: Lawrence Erlbaum Associates.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, *23*, 156–166.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2429.

Henson, R. N. A., Shallice, T., & Dolan, R. J. (1999). Right prefrontal cortex and episodic memory retrieval: a functional mri test of the monitoring hypothesis. *Brain*, *122*, 1367–1381.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*, 75–116.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923.

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*, 85–98.

Howard, M. W., Kahana, M. J., & Wingfield, A. (submitted). Modeling the associative deficit with aging in the temporal context model. *Psychonomic Bulletin and Review*.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging*. Massachusetts: Sinauer Associates, Inc.

Ishai, A., Ungerleider, L. G., & Haxby, J. (2000). Distributed neural systems for the generation of visual images. *Neuron*, *28*, 979–990.

Jay, T. M., & Witter, M. P. (1991). Distribution of hippocampal ca1 and subicular efferents in the prefrontal cortex of the rat studied by means of anterograde transport of phaseolus vulgaris-leucoagglutinin. *The Journal of Comparative Neurology*, *313*, 574–586.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, *24*, 103–109.

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journal of Gerontology*, *60*.

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and

lag-recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 530–540.

Kahn, I., Davachi, L., & Wagner, A. D. (2004). Functional-neuroanatomic correlates of recollection: Implications for models of recognition memory. *Journal of Neuroscience*, *24*, 4172–4180.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302.

Koppenaal, L., & Glanzer, M. (1990). An examination of the continuous distractor task and the "long-term recency effect". *Memory and Cognition*, *18*, 183–195.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Laming, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology*, *34*, 419–426.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. R. (1998). Efficient backprop. *Lecture Notes in Computer Science*, *1524*, 9–50.

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, *23*, 3963–3971.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fmri signal. *Nature*, *412*, 150–157.

MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding, & C. M. MacLeod (Eds.), *Intentional forgetting: interdisciplinary approaches* (Chap. 1). New Jersey: Lawrence Erlbaum Associates.

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*, 654–665.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefontal cortex of the macaque. *Journal of Neuroscience*, *16*, 5154.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *5*, 145–175.

Morris, R., Pandya, D. N., & Petrides, M. (1999). Fiber system linking the mid-dorsolateral frontal cortex with the retrosplenial/presubicular region in the rhesus monkey. *The Journal of Comparative Neurology*, *407*, 183–192.

Moscovitch, M., & Winocur, G. (2002). The frontal cortex and working with memory. In D. T. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 188–209). New York: Oxford University Press.

Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of environmental changes on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 882–894.

Murnane, K., & Phelps, M. P. (1994). When does a different environmental context make a difference in recognition? A global activation model. *Memory and Cognition*, *22*, 584–590.

Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 158–172.

Neath, I. (1993). Contextual and distinctive processes and the serial position function. *Journal of Memory and Language*, *32*, 820–840.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *4*, 611–646.

Norman, K. A., & Schacter, D. L. (1996). Implicit memory, explicit memory, and false recollection: A cognitive neuroscience perspective. In L. M. Reder (Ed.), *Implicit memory and metacognition*. Hillsdale, NJ: Erlbaum.

Nyberg, L., Habib, R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences*, *97*, 11120.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, *12*, 1013–1023.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199–1242.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.

Parkin, A. J. (1993). *Memory: phenomena, experiment, and theory*. Cambridge, MA: Blackwell.

Peterson, L. R., & Peterson, M. R. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.

Petrusic, W. M., & Jamieson, D. G. (1978). Differential interpolation effects in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 101–109.

Polyn, S. M., Cohen, J. D., & Norman, K. A. (2004a). Detecting distributed patterns in an fmri study of free recall. *Society for Neuroscience Abstracts*.

Polyn, S. M., Norman, K. A., & Cohen, J. D. (2002). Connectionist modeling of source memory phenomena. *Poster presented at the Society for Neuroscience convention*. Orlando, FL.

Polyn, S. M., Norman, K. A., & Cohen, J. D. (2003, April). Modeling prefrontal and medial temporal contributions to episodic memory. *10th Annual Meeting of the Cognitive Neuroscience Society*.

Polyn, S. M., Nystrom, L. E., Norman, K. A., Haxby, J. V., Gobbini, M. I., & Cohen, J. D. (2004b, June). Using neural network algorithms to investigate distributed patterns of brain activity in fMRI. *Meeting of the Organization for Human Brain Mapping*. Budapest, Hungary.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 55–85). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pucak, M. L., Levitt, J. B., Lund, J. S., & Lewis, D. A. (1996). Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *Journal of Comparative Neurology*, *376*, 614–630.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Rumelhart, D. E. (1996). Backpropagation: The basic theory. In P. Somlensky, M. C. Mozer, & D. E. Rumelhart (Eds.), *Mathematical perspectives on neural networks* (pp. 533–566). New Jersey: Lawrence Erlbaum Associates.

Russchen, F. T., Amaral, D. G., & Price, J. L. (1987). The afferent input to the magnocellular division of the mediodorsal thalamic nucleus in the monkey, macaca fascicularis. *The Journal of Comparative Neuroanatomy*, *256*, 175–210.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1064–1072.

Schacter, D. L. (1987). Memory, amnesia, and frontal lobe dysfunction. *Psychobiology*, *15*, 21–36.

Schacter, D. L., Harbluk, J. L., & McLachlan (1984). Retrieval without recollection: an experimental analysis of source amnesia. *Journal of Verbal Learning and Verbal Behavior*, *23*, 593–611.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 289–318.

Schnider, A. (2001). Spontaneous confabulation, reality monitoring, and the limbic system - a review. *Brain Research Reviews*, *36*, 150–160.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science*, *168*, 1601–1603.

Shimamura, A. P., Jurica, P. J., Mangels, J. A., Gershberg, F. B., & Knight, R. T. (1995). Susceptibility to memory interference effects following frontal lobe damage: Findings from tests of paired-associate learning. *Journal of Cognitive Neuroscience*, *7*(2), 144–152.

Shimamura, A. P., & Squire, L. R. (1991). The relationship between fact and source memory: findings from amnesic patients adn normal subjects. *Psychobiology*, *19*, 1–10.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies, & D. M. Thomson (Eds.), *Memory in context: Context in memory.* (pp. 13–34). Oxford, England: John Wiley & Sons.

Squire, L. R., & Zola-Morgan, S. M. (1991). The medial temporal lobe memory system. *Science*, *253*, 1380–1386.

Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: with applications in physics, biology, chemistry, and engineering*. Mass.: Addison-Wesley.

Thapar, A., & Greene, R. L. (1993). Evidence against a short-term-store account of long-term recency effects. *Memory and Cognition*, *21*, 329–337.

Tulving, E. (1983). *Elements of episodic memory*. Oxford, England: Clarendon Press.

Tulving, E. (2002). Chronesthesia: conscious awareness of structured time. In D. T. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function* (Chap. 20). New York: Oxford University Press.

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory and Cognition*, *30*, 885–892.

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1196–1210.

Watkins, M. J., & Peynircioğlu, Z. F. (1983). Three recency effects at the same time. *Journal of Verbal Learning and Verbal Behavior*, *22*, 375–384.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89–104.

Wheeler, M. A., Stuss, D. T., & Tulving, E. (1995). Frontal lobe damage produces episodic memory impairment. *Journal of the International Neuropsychological Society*, *1*, 525–536.

Wheeler, M. E., & Buckner, R. L. (2003). Functional dissociation among components of remembering: control, perceived oldness, and content. *Journal of Neuroscience*, *23*, 3869–3880.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences*, *97*, 11125.

Wickens, D. D. (1972). Characteristics of word encoding. In A. W. Melton, & E. Martin (Eds.), *Coding processes in human memory* (Chap. 8, pp. 191–215). New York: John Wiley and Sons.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440–455.

Yntema, D. B., & Trask, F. P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior*, *2*, 65–74.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.