WHAT IS PRIMING AND WHY?

CHAD J. MARSOLEK

Open almost any cognitive psychology textbook and you find chapter titles like 'Perception', 'Language', 'Memory', and so on. Often the delineations are more specific (e.g., 'Memory encoding' versus 'Memory retrieval' chapters), but almost always the topic of memory is treated separately from the others. Of course, to some extent, this reflects the organization of different literatures and different groups of researchers in the field, an important goal for a textbook. However, an interesting aspect of this segregated organization of topics is that the brain does not appear to respect it very well, especially with regard to memory. Assuming that memory can be defined as changes in information processing due to previous information processing, and assuming that human memory is subserved by physical changes in the nervous system (e.g., modifications of synaptic efficacy, protein synthesis, etc.), all of the cognitive areas of the human brain appear to be involved in memory per se. That is, all areas instantiate physical changes of the sort that underlie the changes in information processing due to previous information processing (i.e., memory).

In this chapter, I suggest that most cognitive theories of memory are cast in a manner that does not satisfactorily posit both what repetition priming is and why it is that way. Often, theories are cast in terms of abstract entities that are more analogous to computer programming functions than to the functions of brain subsystems. In the abstract approach, theories concern broad functional concepts of the sort reflected by the typical organization of cognitive research topics in textbooks. Alternatively, theories can be cast in terms of neurocomputationally dissociable processing subsystems, the functions that they accomplish, their interactions, and the neurally plausible mechanisms that perform those functions and interactions. In this approach, theories can do more than organize past research findings and generate new questions; they can do so in a manner that highlights how the phenomena stem from independently derived properties and principles of how brains implement memory and how brain-like models simulate memory. In other words, such theories may also explain why mental phenomena exist in the manners that they do and not in some other plausible ways.

Admittedly, explaining why mental phenomena exist as they do is not critically important according to some philosophical and scientific positions. However, we currently are in a position to theorize about memory in a manner that draws direct connections to scientific fields and levels of analysis outside of cognitive psychology, so failure to do so is a choice. In the 1940s and 1950s, Hull's (1943) Principles of behavior was perhaps the most influential text in experimental psychology (at least it was the most cited; see Raja, 1956; Spence, 1952). Only 60 years later, very few researchers hypothesize about 'The symbolic construct $\Delta H_b$ as a function of the number of reinforcements' or 'Primary motivation and
reaction potential’ (chapter titles in that book). This is at least in part because the concepts did not make close connections to other fields and levels, especially the newly developing field of cognitive psychology just a few years after publication of that book. Such connections are needed to help explain why behavioral principles of learning work in one way and not in other ways. Will today’s theories of [insert chapter title of contemporary cognitive text here] likewise fail to make the kind of connections that ultimately explain?

In this chapter, I will contrast what I call theories of cognitive functions against theories of functions of neural processing subsystems. In doing so, I will highlight the fundamental differences in their respective theoretical constructs and degrees of explanatory adequacy. After briefly delineating a memory phenomenon of interest in this book, long-term repetition priming, I will describe a theory of priming that is cast in terms of the architecture, functions, and mechanisms of neural processing subsystems. This theory will highlight the virtues of increased explanatory adequacy through mutually reinforcing constraints from qualitatively different kinds of computational, implementational, and behavioral evidence. Finally, I will exemplify theory development within this approach with a brief description of two additional neural subsystems relevant to visual-form recognition and priming.

Cognitive functions versus functions of neural subsystems

Cognitive phenomena can be understood in different ways through different kinds of theories. The following two theoretical approaches are distinguished in large part by the theoretical constructs that are used, but not by their level of explanation/analysis.

Theories of cognitive functions

In early cognitive psychology, the serial digital computer was extremely useful as a model of how to concretely conceptualize the operations of the human mind. Such a model was essential, especially when attempting to overcome the ‘anti-mind’ influence of behaviorism on psychology. It offered the notion that the mind could be thought of as analogous to computer software, a tractable and coherent entity, with the underlying brain being analogous to computer hardware, a quite distinct entity from the software.

This model led to what I will call theories of cognitive functions. Very often, different phenomena of cognition were understood as being accomplished by different ‘programs’ or mathematical functions, not unlike how distinct pieces of code are used as distinct sets of instructions for computers. Cognitive theorists built different theories for cognitive phenomena with seemingly different behavioral properties. The different phenomena typically had their own real world names in natural languages (perception, language, memory, etc.).

Theories in this approach did not necessarily have an explicit commitment to the serial digital computer as a model of the mind. Instead, perhaps the most critical characteristic of theories of cognitive functions was that the theoretical constructs were fairly abstract entities (in line with a kind of implicit commitment to the computer model). Whether the theories took the form of flowcharts of processing with only pseudo-code specification of algorithms or the form of more detailed specifications in operational computer code, the theoretical constructs were abstract in that they were functional entities that did not necessarily resemble qualitative aspects of how the underlying machinery implemented them. For example, the breakdown of different functional entities (e.g., two distinct subprograms of a common program, or two distinct mathematical functions) did not necessarily reflect a spatial separation of the underlying hardware locations. Furthermore, any two hypothesized processes could take place in the same underlying hardware. The breakdowns among different theoretical constructs and their operating characteristics reflected only fairly abstract theoretical entities.

Theories of functions of neural processing subsystems

As it became clear that the human brain does not operate like the hardware of a serial digital computer, however, different approaches to understanding cognition developed. An alternative to theories of cognitive functions is an approach that I will call theories of the functions of neural processing subsystems. In this approach, the computer model of the mind is largely replaced by the brain and its neural computations as something like a ‘model’ of the mind. Perhaps the most critical characteristic of theories in this approach is that the theoretical constructs are less abstract and more closely tied to qualitative aspects of how the underlying machinery implements mental phenomena in neurocomputationally plausible ways. For example, the breakdown of different functions in this approach reflect the breakdown of anatomically separable brain areas responsible for neurocomputationally distinct processes, and the hypothesized processes reflect brain-like information processing (e.g., making use of distributed representations, parallel processing, etc.). The breakdowns among different theoretical constructs and their operating characteristics reflect decidedly less abstract theoretical entities compared with theories of cognitive functions.

It is important to note that functions of neural processing subsystems are cast at the same 'functional' level of explanation (i.e., level of analysis) at which theories of cognitive functions are cast (see Figure 3.1). Within Marr's (1982) framework, both kinds of theories are cast at a 'representation and algorithm' level, in which functions of systems are hypothesized in terms of representations of inputs, representations of outputs, and transformations of one to the other in the system. Neither kind of theory is cast in terms of implementational-level information alone (the underlying physical hardware information per se), nor in terms of the computational-level theory alone (the goal for the output of the system, what is available in the input to help achieve the goal, useful strategies for achieving the goal given the available input, etc.; see Marr, 1982). Although constraints from the implementational and computational levels can help to delineate and specify functions of neural processing subsystems, the theories of those functions are cast at the middle, functional level of explanation, one in which mental representations and processes per se are hypothesized, just like in theories of cognitive functions. The difference between the two kinds of theories is one of abstractness of theory within the same level of explanation (for related discussion, see Broadbent, 1985; Rumelhart and McClelland, 1985).

Long-term repetition priming

The focus of this chapter is on the cognitive phenomenon of long term repetition priming (cf. Cofer, 1967). According to most usages, this term refers to a facilitation in processing a stimulus due to recent processing of that stimulus or a highly similar one (for reviews, see
Theories of long-term repetition priming

Theories of priming: cognitive function

Most contemporary theories of repetition priming and implicit memory are theories of cognitive function. This is relatively easy to discern in fairly abstract 'processing' theories of priming (e.g., Graf and Ryan, 1990; Jacoby, 1991; Roediger, 1990), whether or not they appeal to general principles for predicting memory performance (Morton et al., 1977). It is also relatively easy to discern in highly abstract cognitive models of priming (e.g., activation in the logogen model, Morton, 1968; bias in the counter model; Ratcliff and McKoon, 1997) and in abstract mathematical models of priming (Wagemakers et al., this volume) within the global memory model approach (for review, see Clark and Goodland, 1996). But, perhaps surprisingly, many 'systems' theories of priming should also be viewed as theories of cognitive function (e.g., Squire, 1994; Schacter and Tulving, 1994), at least when the proposed theoretical constructs (e.g., branches in a broad taxonomy of memory) do not correspond directly to the breakdown of neurocomputationally dissociable brain areas and do not propose distinct brain-like processes in some concrete detail. This point will be clarified later in the chapter.

Theories of priming that are theories of cognitive function posit fairly abstract explanations that are useful for some aspects of theoretical adequacy but not others. Such theories certainly are useful for providing the descriptive adequacy (cf. Chomsky, 1965) of organizing previous observations in a rational way and generating new questions and new observations, important goals for any scientific theory. However, they tend to provide relatively little explanatory adequacy (cf. Chomsky, 1965), in that they do little to show how the phenomenon stems from properties and principles that are independently motivated from other, different domains (for similar discussion, see Seidenberg, 1993).

Admittedly, some theories of cognitive function are aimed in the direction of providing explanatory adequacy. Many have the virtue of drawing connections to other domains of cognitive function (e.g., perceptual priming occurs due to normal processing and learning in perceptual recognition mechanisms; Bowers and Kouider, this volume; Schacter, 1990). Some also have the virtue of bringing mathematical tractability and other forms of computational constraint to bear on theories of priming (e.g., mathematical models of memory; Wagemakers et al., this volume), and others have the virtue of bringing neuroscientific evidence and other implementational constraints to bear on theories of priming (e.g., priming occurs independently of explicit/declarative memory and perhaps independently of other forms of implicit/non-declarative memory; Squire, 1994; Schacter and Tulving, 1994). These theoretical moves are aimed in the right direction. However, in these theories, the set of additional, independently motivated constraints often does not suffice to explain why priming occurs in that way and also why it does not occur in other plausible ways. Theories with good explanatory adequacy are needed to make this important specification.

Explaining priming

A critical suggestion in this chapter is that the strongest set of independently motivated constraints from multiple domains is a set in which the theoretical constraints from all three levels of explanation of a processing entity (computational, functional, and implementational) mutually reinforce each other. Of course, using many different constraints is beneficial, whether they be multiple independent sources of evidence, methodological tools, or levels of explanation. Any one source, tool, or level has its limitations. But in addition, when the multiple constraints click together to mutually constrain a particular functional-level theory, they may form an additional, emergent constraint, to the extent that an integrated 'whole' constraint is greater than the sum of its 'parts.' A theory adhering to this kind of higher-order constraint not only offers descriptive adequacy but also a way to exclusively distinguish why the phenomenon occurs in that way and not in other plausible ways.

With today's methodological tools and theoretical concepts, this goal is better pursued from the perspective of theorizing about the functions of neural processing subsystems than from the perspective of theorizing about cognitive functions. Theoretical principles...
from biologically plausible computational models (McClelland, 1993; O’Reilly and Munakata, 2000) and neurocomputational tests of these models, combined with neuromaging and neuropsychological evidence from humans and neurophysiological and neuroanatomical evidence from other primates (in addition to cognitive/behavioral experimentation), can constrain theories of the functions of neural processing subsystems to a greater degree than theories of cognitive functions. This can lead to a theory of priming with explanatory adequacy.

Theory of priming: functions of neural processing subsystems

The desired theory of priming will posit a relevant set of neural processing subsystems at various levels of scale, their functions and processing mechanisms, and how they interact to accomplish relatively complex mental processes. Constraints from computational analyses and evidence, implementational evidence, and functional/behavioral experiments will converge to produce hypotheses of priming that attempt to provide explanatory adequacy.

Computational analysis

General architecture of subsystems

Computational reasoning has been offered for the general architecture of subsystems underlying consolidated memories and unconsolidated memories (McClelland et al., 1995). At a big-picture scale, the brain receives perceptual input, implements internal processing, and controls motoric outputs. In the internal processing, it is computationally useful to discover and store the structure that is common to various processing events, because this subsequently provides efficient ways to appropriately process similar but novel inputs. The structure contains the common information that is slowly abstracted from various previous events (e.g., that pianos can produce music), not the distinctive information about particular events (e.g., the distinctive information in an episodic memory of hearing music from a piano). Neurally inspired models with at least partially overlapping (superimposed) representations of knowledge naturally learn this kind of structure (e.g., Knapp and Anderson, 1984; McClelland and Rumelhart, 1985).

Of course, storing information about novel distinctive events is important as well. However, attempts to quickly store distinctive events in models with the superimposed representations that are useful for learning the common structure lead to a very significant problem: catastrophic interference of the new learning on the retention of the old structure (McCloskey and Cohen, 1989). A computationally plausible solution to this problem is buffered storage of the new events in non-overlapping memories (using less superimposed representations) within a separate subsystem (McClelland et al., 1995). This allows the new events to be stored quickly, without the catastrophic interference. Eventually, some of the new information (e.g., a novel musical instrument) can be integrated into the old structure knowledge in the first subsystem, as long as slow learning of the new information is interleaved with reprocessing and ‘relearning’ of the old information (e.g., pianos). In this computational solution (validated through modeling studies reported by McClelland et al., 1995), one subsystem stores the well consolidated common structure and another subsystem stores unconsolidated memories of recently encoded, distinctive information.

Mechanistic causes of priming

Computational reasoning also suggests what the mechanistic causes of priming may be. In the computational theory, ‘relearning’ of the old structure information must occur with some frequency, so it can be interleaved with new learning of novel information and avoid catastrophic interference from the new learning. Interestingly, processing events in everyday life provide just the opportunites needed to relearn the old information involved in these events. Another aspect of the computational theory is that both the relearning trials and the new learning trials must have relatively small effects on the structured knowledge in the consolidated memory subsystem (large changes create catastrophic interference). In neural network models, these changes are small modifications of the weights or connections between processing units in each training trial, for both relearning trials and new learning trials. In this way, the same weights that store the old structure can also integrate new information that updates that structure with information from new experiences.

Therefore, a computationally plausible understanding of priming is that it occurs in the consolidated memory subsystem(s), and it is the memory effect produced by the small changes in this subsystem(s) due to relearning of old information and new integrations of unfamiliar information. For example, assuming visual object recognition requires a subsystem with consolidated visual object knowledge, perceptually recognizing a common object provides the opportunity for small changes to occur in the consolidated memory of that object’s shape, which have the purpose of helping to avoid catastrophic interference. Such changes would have the effect of supporting the facilitation in subsequent recognition of that object that we refer to as repetition priming (for similar modeling approaches to priming and examples, see Becker et al., 1997; Bowers et al., in press; Marsdek and Burgess, 1997; McClelland and Rumelhart, 1985; Rueckl, 1990, this volume).

The effects of priming on subsequent processing

Computational reasoning also suggests important mechanistic effects of priming on the subsequent processing in the relevant subsystem. In network models that perform recognition or categorization of inputs, ‘sparse distributed’ patterns of activation are useful (e.g., O’Reilly and Munakata, 2000; Rolls and Milward, 2000) in the internal representations (i.e., in the activations of the units interposed between the input units and output units of the networks). Sparse distributed activations occur when only a small portion of the units (but more than one unit) is significantly activated (see Figure 3.2). These are patterns of activation that are intermediate between the extremes of local codes (only one unit activated) and densely distributed codes (all units activated), and they often are more useful in computational models than either of the two extreme versions of coding (see Földiák and Young, 1995). In networks that use sparse distributed activations, different units are sensitive to different features or portions of input patterns. Categorizations can be learned through discovering which features are almost always present in the inputs that belong to a particular category (‘presence-diagnostic’ features; e.g., the visual information in an object’s shape that varies little across various viewings), which features are almost always absent for that category (‘absence-diagnostic’ features; e.g., visual information never found in an object’s shape but found in other objects’ shapes), and which features sometimes are and other times are not present for that category (‘non-diagnostic’ features; e.g., visual information specific to one
viewing angle of an object but not another. In this way, a sparse distributed pattern of activation for a category would be strong activation of a small number of units in the internal representation (those representing the presence-diagnostic features for the category) and little or no activation of other units (especially those representing the absence-diagnostic features for the category). With representations like these, the weight changes that occur for relearning of old information and for new learning of unfamiliar information should take place as follows.

First, consider the relearning of familiar information. When a familiar input is recognized in everyday life, an input is accepted by the subsystem and an output representation is produced and sent to subsequent subsystems downstream. The degree to which the representation received by the subsequent subsystems is different from the representations received in the past for that category determines the degree to which the subsequent subsystems can categorize and hence make use of the information effectively (e.g., process the information in ways similar to before). Representations that are different from the norm for that category cannot be categorized or used as effectively as representations that are in line with the norm for that category. This degree of 'usefulness' can serve as feedback from the subsequent subsystems to the first subsystem to guide error-correction learning in the first subsystem (for biologically plausible forms of such error-correction learning, including backpropagation, see Mazzi, et al., 1991; O'Reilly, 1996; O'Reilly and Munakata, 2000). Therefore, when a familiar input is recognized in everyday life, many of the units in the internal representation must be activated in the way that they have been during past successful recognition experiences (otherwise, the input would not be recognized). The connections to these units should not be associated with substantial weight changes, because the units are activated in a manner that is useful for the subsequent subsystems. However, many other units in the internal representation will not be activated in the way that have been in past successful recognition experiences. This is because recognition events since the last encounter with that familiar input will have changed the weights on connections to those units (to 'relearn' those other inputs). The connections to these units should be associated with small weight changes. Such units are responsible for producing a slightly different output representation compared with that produced in most of the past trials for the relevant category.

The important effects on the first subsystem are as follows. Activation of a subset of the units in a subsystem (especially presence-diagnostic units) should contribute to a categorization of the input that is much like the categorizations produced previously by members of that category. For this reason, few error-driven weight changes should occur on the connections to those units. In contrast, activation of another subset of the units (especially units that have not been learned to be absence- or presence-diagnostic) should contribute to a categorization that is not very similar to the categorizations produced previously by members of that category. As a result, error-driven weight changes should occur on the connections to those units in particular. These changes can be of two sorts. Units that contribute to decreases in the usefulness of the output should have associated weight changes that lead to a lesser likelihood of activation by that input (e.g., units that should be learned to be absence-diagnostic). Whereas, units that contribute to increases in the usefulness of the output should have associated weight changes that lead to a greater likelihood of activation by that input (e.g., units that should be learned to be presence-diagnostic). Because of such weight changes, repetition priming for that familiar stimulus would occur (cf. Desimone, 1996; Marsele and Burgund, 1997).

At least two of the changes in the subsystem should be clearly discernible. First, in terms of the activations of individual units in the internal representations, many should not exhibit a discernible change in their likelihoods of activation due to priming. But, others should exhibit a discernible change, some becoming more likely and others becoming less likely to be activated subsequently. Second, in terms of the weight changes that occur in nearly every processing trial, fewer and less in magnitude weight changes should occur after processing a primed stimulus than after processing an unprimed stimulus. This is simply because a primed stimulus enjoys the benefit of recent weight changes that improved its processing, and because of this, subsequent processing of that stimulus leads to an output that should cause fewer and less substantial weight changes compared with a stimulus that has not been primed recently. These are critical hypotheses that will be revisited below.

Now consider the learning of new information: First, note that the new information considered at this point is not complex new event information (e.g., a new associative memory of the episode of hearing music from a piano or from a novel musical instrument); that would be the kind of information stored in the unconsolidated memory subsystem. Instead, the new information considered here is the simpler information relevant to a particular consolidated memory subsystem (e.g., the shape of a novel musical instrument for a visual-form recognition system). Almost by definition, there is no differentiation between diagnostic and non-diagnostic information in an unfamiliar input, at least not until sufficient learning enables the diagnostic information to be abstracted from multiple inputs that should be categorized together (e.g., Knapp and Anderson, 1984; McClelland and Rumelhart, 1985). Instead, the units that happen to be activated and non-activated by an unfamiliar input are the initial candidates for representing the truly presence-diagnostic and the truly absence-diagnostic information for the relevant category, respectively (cf. Marsele and Burgund, 1997; Marsele et al., 1996).

The effects of processing an unfamiliar input should be different from the effects of processing a familiar input. The first presentation of an unfamiliar input should elicit activation of units that happen to be sensitive to the features in that input. When this input is initially discerned to be novel (i.e., the output representation is substantially
different from all past output representations), the weight changes should not be substantial. This is because no ‘error signal’ can be assessed in the subsequent subsystems; no pre-existing representations exist to allow a comparison between the output produced by the first subsystem against outputs produced in past trials. However, a new representation of that input may be stored quickly in the unconsolidated memory subsystem, and if so, subsequent processing of that input can lead to the kind of feedback from subsequent subsystems that signals weight changes in the first subsystem. Through this kind of process, the first subsystem eventually is able to abstract the diagnostic information for the relevant category via subsequent learning trials with additional, different inputs belonging to that category. Alternatively, the input may not be initially discerned to be novel, but instead may be misrecognized as belonging to the most similar familiar category. In this case, the first presentation of the ‘unfamiliar’ input should lead to the kind of repetition priming effect observed for familiar inputs (weights are adjusted—mistakenly in this case—on the connections to a subset of units in the first subsystem, as described above).

Thus, when unfamiliar inputs are discerned to be novel (not misrecognized as familiar), an important change in the subsystem should be clearly discernible. Weight changes should not be substantial for unprimed stimuli (stimuli never previously processed), but should be substantial for primed stimuli (stimuli previously processed once before). Thus, weight changes should be larger in number and magnitude for primed than for unprimed unfamiliar stimuli. This contrasts with how weight changes should be smaller in number and magnitude for primed than for unprimed familiar stimuli. This is a critical hypothesis that will be revisited below.

It is important to note that this understanding of priming stems from general computational constraints on mechanisms that could support human learning and memory. These are properties of learning and memory that are derived independently of behavioral or neuroscientific experimentation with repetition priming per se. Thus far, we have only considered computational constraints. To the extent that such constraints dovetail with independently derived evidence from other levels of explanation—in a mutually reinforcing way—the developing theory may have good explanatory adequacy. The next section summarizes implementational evidence of how the primate brain underlies learning and memory that integrates well with the computational reasoning.

Implementational analysis

General architecture of subsystems Various neuroscientific findings are in line with the general architecture of memory subsystems hypothesized above. The hippocampal formation (and related areas) underlie unconsolidated memory—off-line from the neocortical areas underlying consolidated memory—according to evidence of selective impairment of unconsolidated memory in human amnesia (e.g., Milner et al., 1968), related animal models of amnesia (e.g., Zola-Morgan and Squire, 1990), and patterns of activation in neuroimaging studies with humans (e.g., Schacter et al., 1996). In contrast, neocortical areas underlie consolidated memory storage (e.g., Ungerleider, 1995) as well as priming effects, as evidenced by intact priming following medial temporal damage (for familiar information, e.g., Warrington and Weiskrantz, 1974, and for unfamiliar information under some conditions, e.g., Bowers and Schacter, 1993) and by neural changes during expression of priming in human neuroimaging studies (e.g., Buckner et al., 1995; Squire et al., 1992). A neuropsychological double dissociation between unconsolidated hippocampal memory and neocortical priming effects can be observed (Gabrieli et al., 1995). Moreover, properties of the organization of neocortical areas into perceptual and associative/conceptual areas reflects properties of priming effects. Priming of perceptual and conceptual information can be distinguished (e.g., Schacter, 1992), and neuroactivation patterns in human neocortex reflect this distinction (e.g., Blaxton et al., 1996; Gabrieli et al., 1996).

Mechanistic causes of priming Neuroscientific evidence also integrates well with the hypothesized computational mechanism of priming. The hypothesis that priming is caused by small changes to the ‘weights’ on connections between processing ‘units’ fits well with evidence that modifications of synaptic efficacy occur between neurons in neocortex following recent activation. Long-term potentiation (LTP) and long-term depression (LTD) are synaptic modification effects that have been observed in cells from many areas of neocortex (for LTP: Artola and Singer, 1987; Komatsu et al., 1988; for LTD: Artola et al., 1990; Kirkwood and Bear, 1994), including the visual cortical cells that underlie shape recognition in human cortex (e.g., Chen et al., 1996). These synaptic changes very likely are involved in learning per se in the neocortex. Recent learning can be shown to induce natural synaptic changes in neocortex that affect subsequent measures of LTP and LTD saturation when LTP is electrically induced in slice preparations (Rodić-Pedotti et al., 2000). Furthermore, the properties of these synaptic changes are consistent with biologically plausible versions of error-correction learning in neural network models (O'Reilly and Munakata, 2000).

The effects of priming on subsequent processing Neuroscientific evidence also fits well with the hypothesized effects of priming on subsequent unit activations and on subsequent weight changes in computational models. First, sparse distributed coding is evidenced in neocortex (Vinje and Gallant, 2000), including in the temporal cortical neurons underlying visual shape recognition (Rolls and Tovee, 1995; Young and Yamane, 1992). These neurons are sensitive to visual shape, but it is often time-consuming to find which whole shape maximally activates a particular neuron (e.g., Gross et al., 1972), consistent with sparse distributed coding. Although the coding is relatively sparse, the representations of shape are at least partially superimposed (partially distributed). Different neurons are maximally sensitive to different critical features of larger whole objects (Tanaka, 1993). Such features are simple enough that they are present in multiple whole input shapes (indicative of overlapping, superimposed representations) but also complex enough that they are not present in very many whole input shapes (indicative of relatively sparse distributed activations). In addition, projection of feedback information from subsequently activated subsystems to previously activated subsystems occurs in all anatomiically studied areas of visual neocortex (Felleman and Van Essen, 1991; Van Essen and DeYoe, 1995). Such feedback is the sort needed for the usefulness of categorizations in subsequently activated subsystems to influence changes in the previously activated categorization subsystems. Most important for present purposes, according to a series of
studies (for review, see Desimone et al., 1995), the effects of repeated presentations of shapes on neuronal activations in temporal cortex fit well with the predicted computational effects of visual priming, both when the stimuli are familiar and when they are unfamiliar.

When familiar shapes are used in repetition studies (Miller et al., 1991, 1993; Miller and Desimone, 1994), some cells, but not all, are affected by previous activation. About half of the cells are reactivated by repeated presentations of their ‘preferred’ stimuli to the same high level as during the preceding presentations. This is in line with the hypothesis that, when relearning old information, weight changes should not appreciably affect subsequent activations of many of the units (presence-diagnostic units). In contrast, the other half of the cells exhibit different levels of activation between the preceding and repeated presentations. Of the cells that exhibit a change, the majority show repetition suppression (see also, Badis and Rolls, 1987; Brown et al., 1987), in that lower activation is elicited by the repeated presentations. A minority show repetition enhancement effects instead. This is in line with the hypothesis that, when relearning old information, weight changes should have the effect of making some units less likely of being activated by a primed stimulus but other units more likely of being activated (units that are not yet learned to be absence- or presence-diagnostic). Moreover, several of the properties of such cellular repetition effects mirror behavioral properties of repetition priming for familiar shapes in humans (Desimone, 1996; Wiggs and Martin, 1998). For example, changes in object size or location do not substantially influence behavioral priming effects (Biederman and Cooper, 1992; Cooper et al., 1992; but for evidence that task demands can alter this pattern, see Srinivas, 1996), and size or location changes do not influence cellular repetition effects appreciably (Lueschow et al., 1994).

In addition, human neuroimaging evidence is in line with the computational predictions. When sets of familiar shapes are used, less activation typically is observed in occipital-temporal visual cortex for primed than for unprimed stimuli (for reviews, see Schacter and Badgayan, 2001; Schacter and Buckner, 1998; Wiggs and Martin, 1998). In addition, in a time-extended object-identification task using functional magnetic resonance imaging (fMRI), the deactivation associated with priming has been shown to occur after the test object has been recognized but not before (James et al., 2000). This is in line with the computational hypothesis that overall synaptic modifications should be less for primed than for unprimed familiar stimuli, in an effect that occurs after recognition of a stimulus and subsequent ‘error-correction’ computations.

A different finding is obtained with unfamiliar shapes. When a set of novel shapes is introduced after a set of familiar shapes has been presented many times previously (Li et al., 1993), the cells that previously exhibited repetition suppression effects for the familiar stimuli typically are activated by the subsequent unfamiliar stimuli. This is in line with the hypothesis that, when learning new stimuli, the information in the input should produce activation in many of the units, because nearly all of the information in an unfamiliar stimulus is ‘diagnostic’ of its category (until truly diagnostic information can be gleaned from additional learning trials). In addition, human neuroimaging evidence fits this pattern. Unlike with familiar shapes, when unfamiliar shapes are presented, greater activation is observed for primed than for unprimed stimuli (Henson et al., 2000; see also Schacter et al., 1995). This is in line with the computational hypothesis that new learning of unfamiliar stimuli is distinguished by weight changes that should be larger in number and magnitude for primed than for unprimed stimuli.

It is important to note that much of this implementational evidence provides independent constraints on understanding priming. With the exception of neuroimaging studies originally designed in part to investigate priming per se, the implementational evidence for the developing theory of priming stems from properties of the brain that were derived independently of interest in priming effects and independently of the computational reasoning for what priming may be. Nevertheless, both the implementational and implementational constraints converge on a distinct functional-level understanding of repetition priming.

Functional theory

The theory of priming, cast in terms of functions of neural processing subsystems, is as follows. Neocortical subsystems store well consolidated knowledge. Everyday cognitive processing (such as perceptual recognition of objects) makes use of well-established structure in the knowledge in these subsystems. Priming occurs when processing of a stimulus causes small changes to the knowledge stored in these neocortical subsystems (to help avoid catastrophic interference), which has the effect of facilitating subsequent processing of the primed stimuli. The mechanism involves small structural changes to representations (different for familiar versus unfamiliar stimuli) that are analogous to weight modifications in neural network models and are actually instantiated by synaptic modifications that influence both subsequent cellular activation and subsequent weight changes.

Why?

This theory may be preferred over typical theories of cognitive function because it has stronger explanatory adequacy. It is true that some theories of cognitive function make effective use of computational or implementational constraints in addition to behavioral findings. But, unless they satisfy constraints from as many independently derived properties of learning and memory, from qualitatively different levels of explanation that all converge in a mutually reinforcing way, it can be argued that they offer a smaller degree of explanatory adequacy. In other words, such theories do a less effective job at explaining why priming works in one way and not in the other plausible ways.

For example, much neural evidence indicates that anatomically separate areas underlie priming versus learning of new episodes and explicit memory (e.g., Squire, 1994; Schacter and Tulving, 1994). But, why hypothesize that the neural evidence necessarily indicates that different subsystems underlie these forms of memory when a common memory system could account for the neuroscientific results (cf. Noshofsky and Zaki, 1998; Palmeri and Flaherty, 1998)? Independently derived computational constraints (e.g., that interrelated learning of old and new information is needed to avoid catastrophic interference in the superimposed representations of consolidated memory) helps to choose between these two plausible alternatives.

In addition, many processing theories of priming are supported by findings that memory performance increases when the type of processing at encoding matches the type of processing at test (e.g., Graf and Radvan, 1990; Roediger, 1990). So why hypothesize that such a principle, as it applies to repetition priming, necessarily reflects mechanistic changes in a
Abstract and specific visual-form recognition and priming

In this section, two additional subsystems are briefly described to exemplify theory development within the approach of theorizing about functions of neural subsystems. Stemming from the mechanistic theory of priming above, two critical subsystems of consolidated memory in the neocortex may be an abstract-category subsystem and a specific exemplar subsystem. Both accomplish visual-form recognition and priming, but in different ways to subserve contradictory transformations and goals.

Computational analysis

Fundamental to visual-form recognition is the ability to recognize the abstract category to which an input shape corresponds (e.g., cup versus pen, etc.) as well as the ability to recognize the specific exemplar to which that same input shape corresponds (e.g., an individual pen). Post-visual feedback can help a visual subsystem to learn that multiple input shapes (even dissimilar ones; e.g., an upright piano and a grand piano) should be categorized together because they are associated with the same post-visual information.

Post-visual feedback also can help to learn that multiple input shapes (even similar ones; e.g., two highly similar upright pianos) should be distinguished because they correspond to different individual object entities in the world. Interestingly, mapping an input shape to its category representation and mapping that input to its exemplar representation involve contradictory computations when real world stimuli are considered.

Figure 3.3 helps to convey the computational analysis that leads to the theory of a two subsystem architecture. Object recognition can be conceptualized as instantiating a mapping from points in input space (retinotopically-mapped input representations for a visual-form recognition subsystem) to points in a long-term memory space (output representations from a visual-form recognition subsystem). First, dissimilar exemplars in a category reside in relatively distant points in input space, and they are mapped together for category recognition versus apart for exemplar recognition (Figure 3.3A). Such mappings are not contradictory; they can take place effectively in a common neural network model (Marsalek, 1992; Marsalek and Burgund, 1997). Second, similar exemplars in a category reside in relatively nearby points in input space, and they are mapped together for category recognition versus apart for exemplar recognition (Figure 3.3B). Such mappings also are not contradictory; they can take place effectively in a common neural network model (e.g., Hummel and Stankiewicz, 1998; Knapp and Anderson, 1984; Marsalek, 1992; Marsalek and Burgund, 1997; McClelland and Rumelhart, 1985). However, contradictory mapping solutions are demanded when categories contain both dissimilar exemplars and similar exemplars (as in most real world visual-form categories; e.g., pianos). Assuming a common internal representation for the mappings, the transformations useful for bringing together dissimilar exemplars contradict the transformations useful for separating similar exemplars (Figure 3.3C; Marsalek, 1994; Marsalek and Burgund, 1997). A computationally useful solution is to separate the mappings across different sets of weights and internal representations (i.e., implement separate, parallel subnetworks for the two mappings; see Figure 3.3D). Note that the same argument applies to learning of visual word-form categories and exemplars (e.g., the same word printed in different letter cases or fonts).

The internal representations that are most useful for abstract-category and specific-exemplar mappings may be qualitatively different, suggesting important mechanistic differences between the two subsystems. Both processors should utilize sparse distributed activations, of the sort hypothesized above, but the two may differ in the degree of sparseness that is most useful. For categorizing dissimilar exemplars in a category recognition subsystem, the presence-discriminative features necessarily correspond to a small number of relatively simple features of whole input images, because little visual information is common to the dissimilar exemplars. Hence, very sparse distributed activations are useful in the internal representations. In contrast, for distinguishing similar exemplars in an exemplar subsystem, the diagnostic information for an exemplar corresponds to a large number of relatively complex features of whole input images, because so much visual information is common to the similar exemplars. Hence, less sparse distributed activations are useful in the internal representations. Indeed, a large number of units, each sensitive in different ways to information close to the whole of an input form, can be drawn to represent extremely specific information (e.g., Ballard, 1986; Hinton et al., 1986), as necessary for very fine-grained exemplar recognition (Marsalek and Burgund, 1997). Because a single,
Figure 3.3 Abstract-category mappings and specific-exemplar mappings are not contradictory when categories contain either dissimilar exemplars (A) or similar exemplars (B), but they are contradictory when categories contain both dissimilar and similar exemplars (C). A computationally useful solution to the latter problem is to implement separate subnetworks for the two mappings (D). See text for explanation.

Unified systems cannot represent both very sparse and not very sparse distributed activations, separate subsystems are needed to accomplish both effectively. In line with these computational analyses, neural network modeling studies indicate that separate subnetworks accomplish abstract-category and specific-exemplar recognition (and priming) of objects in ways that mimic human performance in experiments with the same stimuli and the same priming procedures (Marsalek, 1999), but unified networks do not (Marsalek et al., 1999). Also, the two subnetworks discover mapping solutions that make use of very sparse versus less sparse distributed representations, respectively (Marsalek et al., 1999).

Implementational analysis

Neuroscientific findings integrate well with the hypothesis that abstract-category and specific-exemplar visual-form recognition systems are dissociable and operate in parallel. Hemispheric asymmetries in normal human observers indicate that the pattern of priming that is characteristic of an abstract-category subsystem and the pattern of priming that is characteristic of a specific-exemplar subsystem are neurally dissociable (Marsalek, 1999; Marsalek and Bourguil, this volume). In divided-visual-field experiments, participants name objects presented in the left or right visual field during a test phase. This occurs after they have viewed centrally presented same-exemplar objects, different-exemplar objects, and printed words that name other objects during an initial encoding phase. Priming is abstract yet visual (equivalent same- and different-exemplar primed performance, with both greater than word-primed performance) when test objects are presented directly to the left cerebral hemisphere. But, priming is exemplar-specific (greater same- than different-exemplar primed performance, with the different-exemplar primed performance being equal to word-primed performance) when test objects are presented directly to the right cerebral hemisphere. The evidence of abstract priming without accompanying specific priming in left-hemisphere presentations and specific priming without accompanying abstract priming in right-hemisphere presentations suggests that the two processors operate in parallel rather than in sequence. For a review of similar findings (including word-form priming effects) in normal participants as well as relevant results from brain-damaged patients and neuropsychological syndromes, see Marsalek and Bourguil (1997).

Neuroimaging evidence also is in line with the proposed neural system architecture. Deactivations associated with object priming in a rapid event-related fMRI study (using the same stimuli as in the divided-visual-field and neural-network modeling studies) indicate different neural locations for abstract-category priming and specific-exemplar priming, in both the left and right hemispheres (Marsalek et al., 2001; for similar results, see also, Bockner et al., 1998; Koutstaal et al., 2001). Also, in line with different levels of sparse distributed coding, the abstract and specific areas differed in relative sizes and locations (Marsalek et al., 2001).

Functional theory

The resulting functional theory is that two dissociable visual-form subsystems operate in parallel to accomplish abstract-category recognition and priming versus specific exemplar recognition and priming. Both subsystems use sparse distributed activations in their internal representations, but they use different relative levels of sparseness to accomplish contradictory processes and goals. Both appear to operate in each hemisphere, but the abstract subsystem operates more effectively than the specific subsystem in the left hemisphere, and the specific subsystem operates more effectively than the abstract subsystem in the right hemisphere.

Of course, in typical real-world situations, human observers recognize inputs in the center of vision with multiple fixations, such that the visual information is projected directly to both hemispheres. An interesting aspect of the subsystems theory is that task and
stimulus demands are critical for hypothesizing whether one subsystem or the other will win the race to guide post-visual processing, and hence which one will be exhibited in performance following central presentations. For example, presenting visually degraded stimuli (which selectively affects very sparse coding mechanisms) and manipulating task demands so that recognition requires processing of relatively specific information both lead to decreased contributions from an abstract-category subsystem relative to a specific- exemplar subsystem in priming experiments (Burgund and Marsolek, 1997; Marsolek, 1999; Marsolek and Burgund, this volume; Marsolek and Hudson, 1999).

This theory also leads to interesting explanations for previously puzzling inconsistencies in the perception and memory literatures. For example, some experimental evidence indicates that visual object priming is not specific to particular viewing angles of the same object (e.g., Biederman and Gerhardstein, 1993), whereas other evidence indicates that it is (e.g., Tarr, 1995). This discrepancy may be due in part to tapping different abstract and specific subsystems, as viewpoint-invariant priming is observed in left-hemisphere presentations and viewpoint-specific priming is observed in right-hemisphere presentations (Burgund and Marsolek, 2000). Similarly, experimental results both for and against abstractionist and episodic-specific effects in visual word-form recognition may be explained in part through tapping different subsystems underlying abstract and specific word-form priming effects (Marsolek, 2001). Furthermore, by the theory above, priming of unfamiliar visual-form information may occur in neocortical subsystems, and such priming effects may occur in addition to new memories for the unfamiliar information in an unconsolidated memory subsystem. A recent study explores how these two sources of memory for unfamiliar objects can differ yet both contribute to memory expression (Burgund and Marsolek, 2001).

Conclusion

Why does long-term repetition priming occur? A theory with explanatory adequacy is needed to provide a satisfactory answer, one that explains why it occurs in one way and not in other plausible ways. Such a theory must rationally organize previous behavioral observations of priming well enough to account for them and well enough to generate interesting new questions and research. But, such a theory also must indicate how the phenomenon stems from independently motivated properties and principles from multiple domains and levels of explanation, all of which mutually constrain each other. A theory of long-term repetition priming with explanatory adequacy can be offered, stemming from hypotheses about the architecture, functions, and processing mechanisms of neural processing subsystems and their interactions. This theory does not fit very cleanly into only one or another of the typical topics in cognitive psychology textbooks. For example, a neocortical visual-form recognition subsystem plays at least some role in visual perception (shape recognition), some aspects of language (word recognition and reading), some aspects of memory (consolidated memory of visual shape, priming of familiar and unfamiliar visual shapes), some aspects of categorization (for shape information), and other areas of cognition. It may be interesting that a cost to pay for developing such a theory with explanatory adequacy is that it does not easily adhere to the traditional organization of cognitive research.

Acknowledgments

Writing of this chapter was supported by the National Institute of Mental Health, Grant MH60442. I sincerely thank Hedy Amiri, David Andreason, Jeff Bowers, Lisa Elo, and Carmen Westerberg for helpful comments on an earlier draft, and Chris Axelson for insightful discussion.

References


